



**THEORY AND PRACTICE
OF PSYCHOLOGICAL
TESTING** REVISED EDITION



FRANK S. FREEMAN

PREFACE TO THE REVISED EDITION

THIS revised edition, while essentially the same in form as the original one, has been improved and brought up to date where possible. The following special features might be noted. In this edition there is a fuller discussion of test standardization, particularly as regards methods of estimating reliability and validity. Some of the more significant recently published tests have been included. The treatment of projective techniques has been extended in such a manner as to be particularly useful to students who are not specializing in clinical psychology. Also, the discussion of tests of specific aptitudes has been extended. Throughout, an effort was made to incorporate in discussions and evaluations the results of representative researches that have appeared since the publication of the first edition of this book. One other point in particular, should be noted here: namely, that considerably more attention and emphasis are given in this edition to psychological analysis of functions being tested by each of the several types of measuring devices. This aspect of the subject was not neglected in the first edition; but it has been enlarged in this revision. This is not to say that factorial analysis is disregarded; it signifies, however, that the value of such analysis rests basically upon the psychological insights of the test builder at the outset.

Throughout this edition, more so than in the first, emphasis has been placed upon the necessity of interpreting test results in the light of the psychological principles involved, of the statistical bases in test construction, and of an understanding of developmental and behavioral principles. It is the hope of many of us in this field of psychology that through such emphasis mechanical use and rule-of-thumb interpretation of tests will be discouraged; while, on the other hand, the importance of general competence in psychology is stressed.

The major deletions are the chapters on "Statistics in Psychological Testing" and on "Applications and Problems" (Chapters 2 and 16, respectively, in the first edition). These chapters have been omitted to make room for essential additions and elaborations germane to the psychological tests themselves. Instructors who found Chapters 2 and 16 of the first edition useful will probably be able to refer their students to that volume for these materials.

January, 1955
Ithaca, New York

F. S. F.

PREFACE TO THE FIRST EDITION

AN EXAMINATION of the volumes by Guy M. Whipple, *Manual of Mental and Physical Tests* * will reveal the changes and developments that have taken place in the field of psychological testing in the last forty years. At the same time, such an examination will also reveal the extent of the indebtedness of modern testing practices to the work of Whipple, his contemporaries, and his predecessors. When Whipple wrote his *Manual*, he did so in order to bring together, for the first time, a comprehensive and balanced description of psychological tests, representing what he called the "simpler" and the "complex" processes.

In the developing scientific field of psychological testing, there is recurrent need for periodic presentation of comprehensive descriptions of tests. This volume is intended to meet such a need. At the same time, however, I have not limited this volume only to descriptions of these psychological instruments.

Shortly after having begun, some years ago, to teach courses in psychological testing, individual differences, and clinical procedures, I was convinced that clinicians and other users and interpreters of test results must have an understanding of the theoretical principles and assumptions upon which tests are constructed. For that reason I have at several points in this volume presented basic theories and principles, independent of any specific test or group of tests; theories and principles which are common to a wide range of tests. In addition, other theoretical principles, assumptions, and problems have been presented, where relevant, in conjunction with the descriptions and evaluations of particular tests or with several belonging to the

* Originally published in 1910. Revised and published in two volumes in 1914 and 1915 (Baltimore, Warwick and York).

same category of instruments. The mere discussion of basic theories, entirely separate and distinct from descriptions and evaluations of specific testing devices, is a relatively barren procedure for anyone except those who are already experienced and qualified in the subject. It is for this reason that I have described many representative tests in detail, both as to psychological and statistical aspects. The student will thus have specific substance to project against and compare with theories previously presented. Knowledge of how to interpret tests is best achieved through combined understanding of theory and familiarity with test content.

Probably no two authors would be in complete agreement as to tests to be included, though their selections would have many in common. I believe, however, that the devices included in this book are representative of the sounder instruments currently available; though in a few instances poor tests have been mentioned or described for the purpose of illustrating a relevant aspect of the subject. The tests have been so grouped, it is hoped, as to prove most useful to those interested in particular types or levels.

Some historical and developmental background of the subject is provided, especially the work of Alfred Binet; for I believe the student achieves a fuller and sounder appreciation of the present status of the science and its applications through a presentation of early work and thinking. For the advanced student, however, the historical and developmental background provided herein should not suffice; he should consult more detailed and comprehensive studies.*

This volume is intended primarily for students who plan to enter professions in which psychological tests are administered and the results interpreted in dealing with adjustment and numerous other psychological problems. Thus it is designed for the use of clinical psychologists, school psychologists (who are also clinicians), guidance counselors, teachers, psychiatrists, pediatricians, social workers, and personnel officers. Without at all minimizing the usefulness of group testing and group studies of psychoeducational, psychosocial and purely psychological problems, the emphasis herein is on individual and clinical interpretation of test findings. While recognizing that an individual does not exist in a social vacuum and that his

* E. g., Jos. Peterson, *Early Conceptions and Tests of Intelligence*, 1925. (Yonkers, N. Y., World Book Co.); E. J. Varon, "The Development of Alfred Binet's Psychology," *Psychological Monographs*, No. 207, 1935.

behavior and performances can be fully understood only in terms of himself and of his environments, the fact is, nevertheless, that in psychological testing the ultimate unit of concern to the examiner usually is the individual subject. After the results of testing are available, in an individual instance, the psychologist may then seek causal and explanatory factors.

While formal preparation in statistics, especially in statistical reasoning, is desirable, even in the case of students beginning the study of psychological testing, it is not possible to provide that preparation in a book such as this. But understanding of the more common statistical indexes, methods, and reasoning is necessary. Hence, a chapter on "Statistics in Psychological Testing" has been included for those students who have not had formal preparation. I am pleased to acknowledge my indebtedness to my colleague, Professor T. A. Ryan, who collaborated by assuming major responsibility for this chapter which is within the area of his major teaching interests.

Psychologists will note, of course, that a large and important segment of psychometric methods and devices has been almost entirely omitted from this book; that is, the psychophysical methods of Weber, Fechner, Müller, Urban, and their successors. The reason for the omission is that psychophysical measurements comprise an area in themselves and are beyond the scope and purpose of this volume.

Any author of a textbook is indebted, of course, to the many scientists and scholars who have preceded him and contributed the materials from which the book is developed. In specific instances of indebtedness, I have acknowledged authors and sources at each point, where the documentation is most useful. In particular, I wish to thank the authors and their publishers who kindly gave permission to reproduce textual, tabular, and graphic materials.

One or more chapters were read in manuscript by Dr. Solomon Machover, Professor Max L. Hutt, and Professor Frederick L. Marcuse. I wish to acknowledge my appreciation of their valuable criticisms and suggestions.

F. S. F.

Cornell University
November 17, 1949



CONTENTS

CHAPTER	PAGE
1. BASIC THEORETICAL PRINCIPLES	1
<i>Objectivity in Administering and Scoring. A Representative Population Sample. Sampling of Traits and Functions. Steps in the Development of a Test. Reliability. Validity.</i>	
2. INTERPRETATION OF TEST SCORES: QUANTITATIVE AND QUALITATIVE	42
<i>An Index of Relative Rank. Psychological Measurement Contrasted with Physical Measurement. Clinical Aspects. Difference between Norms and Standards. Factors in Selecting a Test.</i>	
3. DEFINITIONS AND ANALYSES OF INTELLIGENCE	60
<i>Definitions of Intelligence. Two Comprehensive Definitions. Implications for Test Design and Content. Three "Kinds" of Intelligence. Analyses of Mental Ability. Factor Analysis. Illustrations of Factors. Implications.</i>	
4. THE BINET SCALES	96
<i>Historical Background. The Early Work of Alfred Binet. The 1905 Binet-Simon Scale. The 1908 Binet-Simon Scale. The 1911 Revision of the Binet Scale. Summary.</i>	

CHAPTER	PAGE
5. EARLY REVISIONS OF THE BINET-SIMON SCALE	114
<i>Four Early Revisions. The Stanford Revision of 1916.</i>	
6. THE 1937 REVISION OF THE STANFORD-BINET SCALE	129
<i>Description of the 1937 Scale. Validation. Reliability. Determining Mental Age and Intelligence Quotient. Distribution of IQ's. Suggested Classification of Revised Stanford-Binet IQ's. Analysis of Functions Tested. Types of Items. The Short Scale. Evaluations and Criticisms.</i>	
7. THE WECHSLER SCALES	156
<i>Description of the Wechsler-Bellevue Intelligence Test. Functions Involved in the Subtests. Need for an Adult Scale. Standardization. The Population Sample. Validity. Reliability. Scoring and IQ Calculation. Special Features of the Bellevue Scale. Criticisms and Evaluations. The 1955 Revision of the Bellevue Scale. The Wechsler Intelligence Scale for Children (1949).</i>	
8. INDIVIDUAL PERFORMANCE SCALES	197
<i>The Pintner-Paterson Scale of Performance Tests. The Cornell-Coxe Performance Ability Scale. The Arthur Point Scale of Performance Tests. Revised Arthur Scale: Form II. Other Performance Tests. Functions Tested by Performance Scales. Evaluation of Performance Tests.</i>	
9. SCALES FOR INFANTS AND PRESCHOOL CHILDREN	220
<i>Gesell Developmental Schedules. Minnesota Pre-school Scale. Cattell Developmental and Intelligence Scale. Merrill-Palmer Scale of Mental Tests. Evaluation of Scales for Infants and Preschool Children.</i>	
10. NONVERBAL GROUP SCALES OF MENTAL ABILITY	241
<i>Beginnings. Characteristics of Group Tests of Mental</i>	

Ability. Pintner-Cunningham Primary Test. Chicago Nonverbal Examination. Revised Army Beta Examination. Pintner Nonlanguage Series: Intermediate Test. Nonlanguage Multi-Mental Test. Pattern Perception Test. Progressive Matrices Test. Cattell Culture-Free Test. Goodenough Drawing Test. Davis-Eells Test of General Intelligence. Evaluation of Nonverbal Group Scales.

11. VERBAL AND MIXED GROUP SCALES OF MENTAL ABILITY

270

California Tests of Mental Maturity. Terman-McNemar Test of Mental Ability. Tests of Primary Mental Abilities. Kuhlmann-Anderson Tests (6th Edition). Group Scales for College Freshmen. Army General Classification Test. Miller Analogies Test. Other Group Scales. Evaluation of Group Scales. Uses of Group Scales.

12. APTITUDE TESTS: MECHANICAL AND CLERICAL

306

Definition and Explanation. Tests of Vision and Hearing. Motor and Manual Tests. Tests of Mechanical Aptitude. Tests of Clerical Aptitude. Differential Aptitude Tests. Aptitude Classification Tests.

13. APTITUDE TESTS: FINE ARTS AND PROFESSIONS

336

Tests of Musical Aptitude. Tests of Aptitude in the Graphic Arts. Tests of Aptitude in Medicine. Tests of Aptitude in Law. Tests of Aptitude for Teaching. Tests of Science and Engineering Aptitudes. Interest Inventories. General Evaluation of Aptitude Tests. Steps in an Aptitude Testing Program.

14. TESTS OF EDUCATIONAL ACHIEVEMENT

377

Scope. Purposes. Derived Indexes. Types of Items. Three Representative Batteries. Reading Tests. Arithmetic Tests. Tests at High School and College Levels. Tests of Aptitude in Specific Academic Subjects. Tests of More Complex Educational Objectives. Evaluation of Achievement Tests. Tests of Proficiency.

CHAPTER

PAGE

15.	INTELLIGENCE TESTS AS CLINICAL INSTRUMENTS	403
	<i>Factors Which Affect Test Performance. The Stanford-Binet Scale. The Bellevue Scale. Kent Series of Emergency Scales. A Report Outline.</i>	
16.	TESTS OF MENTAL IMPAIRMENT	432
	<i>The Babcock Test. Tests of Concept Formation. The Hunt-Minnesota Test for Organic Brain Damage. The Bender Visual-Motor Gestalt Test. Evaluation of Tests of Impairment.</i>	
17.	PERSONALITY RATING SCALES	452
	<i>Definition of Personality. Rating Scales: Major Aspects. Representative Rating Scales. Evaluation of Rating Scales.</i>	
18.	PERSONALITY INVENTORIES	466
	<i>Purposes and Types of Inventories. Representative Inventories. Biographical Data Questionnaires. Tests of Attitudes and Values. Opinion Polling. Evaluation of Personality Inventories.</i>	
19.	PROJECTIVE METHODS: THE RORSCHACH AND THE THEMATIC APPERCEPTION TESTS	502
	<i>Definition and Explanation. The Rorschach Test. Thematic Apperception Test.</i>	
20.	PROJECTIVE METHODS: VARIOUS	547
	<i>Word Association Tests. Picture Tests. Verbal Completion Tests. Drawing and Painting. Play. Evaluation of Projective Tests.</i>	
21.	SITUATIONAL TESTS	577
	<i>Sociometric Methods. Tests of "Social Intelligence" and Leadership. Psychodrama. Office of Strategic Services: Assessment Tests. Evaluation of Situational Tests.</i>	
	INDEX	599

I.

BASIC THEORETICAL PRINCIPLES

ALTHOUGH tests of general intelligence, specific aptitudes, personality, and educational achievement are designed and constructed for different purposes, all of them have certain principles and procedures in common; and any combination of these categories of tests might be used in dealing with a specific individual case or in attempting to solve a particular psychological problem. Psychological tests have been used to find answers to a number of psychological questions, both theoretical and practical. But ultimately, and most important, they are intended to contribute to the analysis and description of individuals, and to the evaluation, prediction, and guidance of their behavior and education. The following are the aspects that are common to the several types of psychological instruments and that give them their objectivity: objectivity in administering and scoring; norms based upon a population sampling, scientifically selected, for a particular test; sampling of specified traits or functions, by means of a particular test; incorporation, within a test, of a composite of views of a number of experts; utilization of recognized techniques of test standardization. The tests' objectivity and the standardization process give them a scientific quality which, of course, is absent in an individual's personal estimate of psychological traits and functions. *We define a psychological test as a standardized instrument designed to measure objectively one or more aspects of a total personality, by means of samples of performance or behavior.*

OBJECTIVITY IN ADMINISTERING AND SCORING

Each psychological test is administered under a prescribed set of procedures. These procedures involve preparation of the persons to be tested by means of introductory and explanatory remarks; phrasing and formulation of instructions whereby each part, or each item in some instances, is to be presented; setting time limits, if any; decisions as to when to repeat instructions or to offer encouragement, and when not to do so, as well as when to answer questions asked by a subject, and when not; the use of practice exercises, if any.

The scores and ratings thus derived from the tests are not dependent upon the individual bias or judgment of the particular examiner. For the score of any subject on an objective test is arrived at by the use of a scoring key; or the scoring is otherwise so clearly defined, specified, and illustrated that subjective judgments of individual examiners or scorers do not enter in at all or are reduced to a minimum. Thus an objective test provides a highly uniform means of evaluating the psychological traits or functions being measured; results obtained by one competent examiner are comparable with those obtained by others.

A REPRESENTATIVE POPULATION SAMPLE

Every test is designed and intended for use with a specified *population*, or group. For example, a test of intelligence may be standardized for use with individuals from the age of two years through adulthood (Stanford-Binet, 1937 revision); another for ages eleven to seventeen (Chicago test of Primary Mental Abilities); another primarily for adults (Wechsler-Bellevue). Still others cover different age ranges.

A test of scholastic achievement in a particular school subject or group of subjects may be intended for the first three grades; or for grades eight through twelve; or for college freshmen; or for other grade ranges, depending upon the school subject and the prescribed scope.

Tests of specific aptitudes likewise are designed for specified populations. For example, one test of ability in art is designed for grades seven and above; one test of mechanical aptitude is to be used for ages eight to twenty-one; a law aptitude test is standardized, of course, for

college students and others who are candidates for admission to a law school, regardless of age.

Rating scales, personality inventories, and projective tests are likewise intended for use with a specified segment of the total population. They may be designed for a selected age group; for particular occupations; for given educational levels; for one sex of limited age range; for the diagnosis of clinical cases, or for use with non-clinical populations as well.

In any event, whatever the traits or functions to be measured, whatever the range of ages or school grades, and whether for clinical or non-clinical groups, the test must be standardized upon a group that is a representative sample of the total population for which it is intended. Each test must be constructed by means of actually sampling the performance of an adequate group which has been selected in such a way as to insure its being typical of the population of which it is a part.

Factors to be taken into account in making a population sampling will depend upon the nature and the comprehensiveness of the test under construction. In any instance, the sample should yield unbiased data on the population of which it purports to be representative; and the sample should be large enough to provide statistically valid results for the traits or functions being measured by the test.

This means, of course, that the author of a test must decide at the outset with which group, with what segment of the population his instrument is to be used. Then he must standardize his test upon a population sample that is stratified according to relevant factors; and within each stratum the selection of cases should be adequate in number and of correct proportion in the total.¹

For example, if a psychologist is to construct a test of "general intelligence" for American children, in the primary grades, ranging in age from five to nine years, he will have to take into account the following factors in obtaining his standardization population: age, sex, geographic area, parental occupational level, and type of community (urban, village, farm). The author of the test must decide, also, whether he will standardize his test entirely on a Caucasian population, or whether he will include non-Caucasian elements. If it is to be

¹ On sampling, see M. B. Parten, *Surveys, Polls, and Samples*, New York: Harper, 1950.

the latter, then the racial factor must be taken into account in the stratification of the sample.

Since individuals within a representative sample of children of any given age vary widely in respect to mental abilities, some will reach only the levels of younger age groups while others will attain the levels of older age groups. Thus, to ascertain the developmental level of the retarded it is necessary to extend downward the chronological age of the standardization sample; and, conversely, it is necessary to extend upward the age limit for the superior.

The validity of results obtained with any psychological test will be dependent, in part, upon the adequacy and representativeness of the standardization population.

SAMPLING OF TRAITS AND FUNCTIONS

Any given test measures a limited aspect of the person being examined, though some tests are much more restricted in scope than others. It is essential, therefore, that the test builder define the aspect, or aspects, he proposes to measure. After doing this, he must develop a series of test items that will best sample the traits or functions with which his test is concerned.

In developing a psychological test, it is impossible, and in fact unnecessary, to use an unlimited number of items. It is not necessary to attempt to present the individual being tested (called the "subject" or the "testee") with problems that will ascertain his responses for every conceivable situation involving a given trait or function. It is sufficient to get an adequate sampling of responses in a particular area or range of behavior, the assumption being that the sampling is representative of the whole.

Two kinds of sampling are actually involved in constructing a psychological test. First, the most relevant constituents of the *gross variable* (the broad, comprehensive trait or function) must be selected. Where, for example, the gross variable is "general intelligence," the constituent parts in the test might be: vocabulary, verbal comprehension, arithmetical problems, reasoning with practical problems, verbal and other analogies, perceptual organization, and so forth. Second, the *operational levels* (that is, the actual items) must be selected: *which* arithmetical processes and at what levels, what kinds and which levels of words, what types and range of situations, which perceptual figures?

In following this procedure, psychologists are employing a well-

known and widespread technique. If a chemist wishes to determine the quality of a shipment of milk, he takes small quantities here and there, combines these, and then analyzes a sample of the samples. If an agronomist wishes to analyze a given area of soil, he gathers small amounts from various spots. If a blood test is to be made, a very small quantity, taken from one place, is sufficient and representative of the entire stream. Numerous other illustrations can be found. So, too, with intelligence, specific aptitudes, and school achievement. It has been said that psychological testing may be thought of, figuratively, as sinking shafts here and there within a given range in order to ascertain depth and quality.

Specifically, for present purposes of illustration, *intelligence* may be defined in several ways: (1) capacity to integrate experiences and to meet a new situation by means of appropriate and adaptive responses; (2) capacity to learn; (3) capacity to carry on abstract thinking.² While psychologists differ in regard to which of these three aspects is most important and which they would emphasize, the fact is that most tests of general intelligence probe and sample all three. The following types of items found in various current tests fall under one or more of these definitions, and are constituent parts of the gross variable, general intelligence.

Practical reasoning. "What's the thing for you to do when you have broken something which belongs to someone else?" (From the Revised Stanford-Binet Scale, Form L.)

Definitions of words: i.e., concept formation.

Perceiving similarities and differences between objects: For example: "In what way are wood and coal alike?" "In what way are a baseball and an orange alike and in what way are they different?" (From Revised Stanford-Binet Scale, Form L.): i.e., abstraction and generalization

General information tests: i.e., assimilation and retention of experiences

² Intelligence will be defined and described in a later chapter. There are some psychologists who prefer to discard the term *intelligence* because they believe it is not a function in itself but rather an aggregate of particular aptitudes or, to use a more recent term, of "primary mental abilities." Like many other psychologists, we continue to use the term *intelligence*, which we prefer for two reasons: (1) It has general and meaningful currency now, especially in connection with tests. (2) Even if certain abilities were "primary," intelligence would not be a mere aggregation of these, but rather an integration, in which case intelligence is really something new, different from, and more than its several constituent parts.

Arithmetical reasoning: i.e., reasoning with abstractions

Supplying missing parts to pictures: i.e., perceptual integration

Reproducing geometric figures from memory: i.e., visual imagery and organization

Arranging a series of pictures in logical sequence: i.e., visual perception and reasoning

Perception of color and form design: i.e., visual imagery and recall, analysis and organization

Explanation of absurdities in given pictures: i.e., analysis of visual percepts

Oral solution of practical problems orally presented: i.e., analysis and generalization

Solving problems involving distances and directions (without use of paper and pencil): i.e., spatial orientation

Deriving and giving the meanings from a prose passage: i.e., reasoning with abstractions

Another method of determining the component parts of the gross variable, which in this case we call *general intelligence*, is through "factor analysis," to be discussed more fully in Chapter 3. According to one analysis, there are six such components, relatively independent of one another: ³ facility with numbers (the four fundamental processes), vocabulary (word meaning), space perception (perceiving similarities of and differences between geometric figures), word fluency (controlled word association), reasoning (insight into patterns of letters arranged in series), and memory (immediate recall of discrete verbal materials). Tests have been constructed on the basis of this analysis, items having been devised for each of the six categories.

At present, also, there is a trend among some psychologists toward analysis of the gross variable's component parts into subdivisions; that is, into *elements* of the component parts. For example, the component "reasoning" has been tentatively analyzed into the following four aspects: ⁴

Reasoning I

- a. manipulating symbols
- b. solving problems
- c. defining problems
- d. testing hypotheses

³ L. L. Thurstone and T. G. Thurstone, *The Chicago Test of Primary Mental Abilities*, Chicago: Social Science Research Associates, 1943.

⁴ J. P. Guilford, et al., *A Factor-Analytic Study of Reasoning Abilities*, Los Angeles: University of Southern California, Report Number 1, June 1950.

Reasoning II

- a. seeing rules or principles (induction)
- b. seeing systems
- c. seeing trends
- d. seeing relations (educing relations)
- e. seeing identity of relationships
- f. analyzing forms

Reasoning III

- a. seeing common elements or properties
- b. classifying (in general)
- c. classifying forms
- d. educing correlates

Reasoning IV

- a. drawing inferences (deduction)
- b. syllogistic reasoning

Inspection of these four types of reasoning reveals that they are neither mutually exclusive nor independent of one another. Yet if these and their parts are sufficiently distinct and constitute reasoning in its several aspects, and if reasoning were to be measured according to this scheme, it would be necessary to devise items for each of the four types and for each sub-type.

Specific aptitude, as another example, may be defined as a capacity that indicates the probable degree of successful learning and achievement in a particular and limited type of activity—for example, musical, mechanical, artistic, or linguistic aptitude. A test intended to estimate a person's capacity in each of these must include parts and items sufficient in number and extensive enough in scope to provide an adequate sampling upon which a prediction of subsequent learning and achievement may be based.

The constituent parts of a test of "mechanical aptitude," for instance, might be: knowledge of tools and mechanical devices, skill in assembling parts of a mechanism, perception of spatial relations, manual or digital dexterity, or others found through statistical and psychological analysis to have predictive and selective value.

Personality tests and inventories must also be based upon samplings of the constituent traits that the test author proposes to evaluate. This is true even though personality itself is most difficult to define and though its components are elusive. Thus we have inventories that attempt to measure, among others, degrees of introversion-extroversion, neurotic tendencies, anxiety, hypochondriasis, adjustment to home, adjustment to school, and dominance-submission. Each author of a

personality inventory or of a projective test, in order most adequately to fulfill his purposes, must determine which aspects of personality are to be examined by means of his instrument.

An *educational achievement* test is designed to measure an individual's information in, or skill with, or understanding of—or all three of—a given subject of study taught in school; for example, reading rate and comprehension, arithmetical processes and problem solving, American history, English usage, and so on. In each instance, as in all other types, the scope of the test must be defined, the parts of the gross variable must be determined, and the elements of each part must be represented. Educational achievement tests depend for their validity upon the adequacy with which they sample the subject-matter field for which they are intended.

Tests of general intelligence, of specific aptitudes, and of educational achievement are intended, of course, to indicate the person's status, at the time of examination, in their respective areas; but they are intended, also, for other purposes. Intelligence tests are employed to predict an individual's probable future level of mental development and capacity. Tests of specific aptitude are used to predict probable future learning of and performance in a particular activity or occupation. Results of educational achievement tests are helpful in forecasting the subject's probable future level and quality of learning in the several school areas and in diagnosing specific difficulties and disabilities in basic school subjects. The great importance, therefore, of adequately conceived and satisfactorily standardized instruments is readily apparent.

In respect to determination of an individual's present status, personality inventories and projective tests are like the three other types mentioned above. And while personality tests are used, to some extent, for predicting future behavior and adjustment, their greater significance and usefulness, in relation to one's future status, lies in the fact that their results are very valuable in diagnosing personality problems and as a basis for psychological counseling, or therapy where indicated.

• STEPS IN THE DEVELOPMENT OF A TEST

In devising a test it is necessary, as already explained, first to define that which is to be measured. Tests of intelligence, specific aptitude, school achievement, and personality have already been briefly

defined; and illustrative materials will be presented in later chapters. At this point, several illustrations will suffice to demonstrate how psychological and statistical analysis determine the form and content of a test.

Alfred Binet, the French psychologist of whose work much more will be said later, proposed that the following be tested as components of intelligence: memory, mental images, imagination, attention, comprehension, suggestibility, esthetic appreciation, muscular strength, strength of will, motor skill, and visual judgment. Some of these survived his own experimental investigations and those of other psychologists, while others were rejected as being invalid. Binet suggested the measurement of these processes in the first place because he believed that they differ sufficiently from person to person and that knowledge of their level of development in different persons would permit the psychologist to distinguish one person's general mental development from that of others. Here we have the beginnings of the tests which were to prove so useful in the construction of his and other scales.

Subsequently, psychologists were to identify other and more specific processes which, they held, should be tested. Thus Spearman developed the theory that intelligence is essentially a generalized function (g) and should be measured through a broad sampling of mental activity. E. L. Thorndike, on the other hand, maintained that intelligence consists of a multitude of highly specific processes (unnamed) and that the validity of a test of intelligence will depend upon sampling of these with psychological insight. L. L. Thurstone, among others, developed still a different theory: namely, that what we call "intelligence" is made up of a number of "primary mental abilities," each of which must be sampled, with a "secondary" general factor involved. These theories and other problems concerning the nature of intelligence will be dealt with in Chapter 3; for the present we are concerned only with definitions and analysis of processes as necessary steps in the development of tests.

The Seashore tests in music include the following processes: pitch discrimination, judgment of tone intensity, perception of time, discrimination of tonal timbre, tonal memory, and rhythm discrimination.

In testing knowledge of and skill in language, any or all of the following might be included: grammar, usage, punctuation, vocabulary, reading rate, reading comprehension, visual acuity, eye movements, and auditory acuity.

Regardless of the particular definition of intelligence that commends itself most to a psychologist, and regardless of what analysis into particular processes appears to have greatest value, any psychological test must measure the trait and function in its *manifestations*, in one form or another. The task then is to devise and select items conforming to the definition of the function or trait and to the analysis made of it.

After the original group of items has been devised and selected, they are given a series of try-outs on the groups for which they are intended. The results thus obtained are subjected to established statistical analyses and to scrutiny on the basis of a series of criteria of validity and reliability. As a result of this first analysis and scrutiny, some items are rejected, some are retained, and new ones are added. The new and more highly selected items are again put through the same process of *validation*, resulting in further improvements and refinements of the test under construction. This will be repeated several times before the finished scale emerges. The entire process of try-outs, statistical analysis and evaluation on the basis of accepted criteria is called *test standardization*. Standardization of tests is a topic which will recur with some frequency in this volume. For the present, however, it will suffice to indicate what are the most important aspects involved in the process, in addition to those already discussed.

RELIABILITY

The two essential characteristics of any sound test are *reliability* and *validity*.

The term *reliability* has two closely related but somewhat different connotations in psychological testing. First, it refers to the extent to which a test is *internally consistent*; that is, the extent to which the test scores are subject to or free from such internal defects as will produce errors of measurement due to the quality of the items rather than to the instability of performances of testees themselves. In other words, how accurately is the test measuring at a particular time? Second, *reliability* refers to the extent to which an instrument yields consistent results on testing and retesting. That is, how dependable is it for predictive purposes? Obviously, if a test does not have a high degree of reliability, it can have but limited value, if any, in predicting an individual's future performance or level of development. It is clear that these two aspects of reliability are intimately related; if a test is

not highly reliable on any particular occasion, it can have little predictive value.

Since one of the principal uses of psychological tests is to predict and plan for subsequent development and performance, a high degree of reliability is a *sine qua non* of a sound instrument. Reliability is not, however, an all-or-none proposition; it is a matter of degree. No test is perfectly reliable; the scores obtained on repeated testings are not completely stable, either in terms of internal consistency or of prediction. There are always some errors of measurement, large or small; and it is "normal" for humans to vary in performance, generally within fairly narrow limits, from one occasion to another—to vary, that is, aside from the expected changes that occur as part of the process of growth and development.

Differences among the test scores of individuals in a group are due to: (1) "true," or actual, differences in the trait being measured within those persons being examined, and (2) sources of inaccuracy in the measurement of individuals. These sources of inaccuracy may be inherent defects of the test itself, conditions or "chance" factors operating at the time of testing, or unpredictable fluctuations in the performance of the subjects. Standardization aims to eliminate or reduce inherent defects of the test; the conditions of testing and retesting should be as nearly consistent and optimal as possible; and though minor fluctuations in an individual's performance from day to day or week to week cannot be controlled, the reasons for any major fluctuations must be sought in the individual himself or in some of the environmental forces, if the sources of inaccuracy are to be understood.

The possible sources of variation in performances on a test are many. This aspect of testing will recur frequently in our subsequent discussions; but for the present, the most common of them may be listed as follows:

Actual, or "true," differences among individuals in the *general* traits or *general* abilities being measured.

Specific abilities required in a particular test; or *specific* disabilities in the functions being tested.

Skill in taking tests; being "test wise," or the converse.

The "chance" acquisition of a particular piece of knowledge or information required in a test: e.g., the meaning of an unusual word

such as *ambergris*, or a bit of unusual information such as the name of the author of a little-known work. (These would be poor test items.)

Effects of practice (previous test-taking) or, in some instances, coaching.

Normal or expected fluctuations in performance from time to time.

Personal characteristics of the testee: motivation, health, energy level, emotional status.

Physical conditions under which the test is taken: heat, light, ventilation.

Unpredictable, or "chance" factors: noise, interference, broken pencil, misunderstanding of instructions, etc.

Fortunate guessing of answers.

Test results, ideally, should depend upon the extent to which the test measures the first two of these sources of variation; actually, however, the coefficients of reliability will be adversely affected by the nonsystematic operations of the others.

There are two methods, in general, of expressing the consistency, or dependability, of test results: (1) *absolute reliability*, and (2) *relative reliability*. The first of these is usually stated in terms of the *standard error of measurement*. The second is given, though infrequently, in terms of *analysis of variance*, or, much more commonly, as a *correlation coefficient* indicating the degree to which individuals maintain relatively consistent positions in their group when a single test is administered twice, or when two equivalent forms of a test are applied to all members of a group. This correlation between the two sets of scores is known as the *coefficient of reliability*.

Test-retest reliability. When persons are tested and retested a number of times, they may undergo some change as a result of repeated measurements: e.g., in the form of practice effects, improvement in the skill of taking tests, and in the "set" or attitude toward a test. In estimating reliability, therefore, it is necessary to limit the number of times an individual is tested with the same device. Hence, instead of frequent retesting of the same persons, dependable results for a given psychological instrument are obtained by increasing the number of persons tested rather than by increasing the number of measures of

each person. Therefore, techniques have been devised for evaluating the results obtained with only one or two measurements of the same individuals, namely:

Two equivalent forms of the test are administered and the two sets of scores are correlated.

A single form of the test is administered twice and the two sets of scores are correlated.

The items of a single test are subdivided into two separately scored groups, the two sets of scores being correlated as though they were obtained from two forms or two testings.

Administering two equivalent tests has several disadvantages. The procedure requires more time, of course. The two forms might vary somewhat in content, thus underestimating reliability of either form. The experience of having taken the first test might result in some learning or improved skills. If two forms of a test are to be equivalent, they must be in the same format and each must test a representative sampling of items measuring the same mental processes. Also, the original testing and the retesting should take place within a week or two in order to minimize the influence of intervening factors of developmental and other individual changes.

Administering the identical test twice has some of the disadvantages of using two equivalent forms. It is held by some investigators that recall of answers to specific items of a test is an added disadvantage, when the identical test form is given a second time. Although there can be some recall, it is unlikely that this possibility will be an important consideration, for the number of items in any test is too large for the retention of many. When this method of estimating reliability is used, the interval between testings should be a week or two in order to minimize the effects of whatever recall might be operative.

Split-half reliability. A test cannot have high consistency in retesting, unless each application is relatively free of chance or random errors. In *split-half* testing of reliability, chance and random errors may be assumed to operate equally in both halves.

Calculating reliability by the *split-half* method consists of subdividing the whole test into two parts, presumably equivalent, and then treating the score of each part as though it were a separate form. This method provides, essentially, a measure of the test's internal consistency, assuming an equal level of performance throughout the test by each person. Split-half reliability is a first check upon the usefulness

of a test. It is easily found and saves unnecessary labor that might be spent in following up an internally unsound device. This method tells us if the test is a reliable representation of an individual's traits at a given time. It does not describe completely the reliability of a test which is to be used periodically or for predictive purposes. For periodic and predictive testing, the test-retest method is desirable.

The split-half method of determining reliability may, in some circumstances, yield a coefficient of correlation that is somewhat too high. In calculating reliability, an assumption is that the operations of chance factors are uncorrelated and hence will cancel out one another. But in using the split-half method, both obtained measures are determined at the same sitting and any chance fluctuations due to temporary conditions within testees and to conditions in the external situation will operate in the same direction and thus yield a somewhat higher correlation coefficient than might be found by other methods.

Generally, for split-half reliability, the subdivision is made by taking the odd-numbered items as one part of the test and the even-numbered items as the other. The score is then found for each person, for each of the subdivisions. (This method is referred to as *odd-even* reliability.) Since the correlation coefficient for the two sets of scores derived by this method is based upon subdivisions of the full test, each of which is half the length of the whole, a statistical formula (Spearman-Brown) is used to correct for the reduced lengths of the subdivisions from which the correlated scores have been determined. The reason for this correction is that the score of the whole test, being based upon a larger number of items, is a more adequate sampling of traits or functions and hence reduces the possible effects of chance solutions and accidental errors. The whole test is thus more reliable than its subdivisions; and the correction formula is intended to indicate what the reliability of the entire test would be, based upon what was found with the part scores.

An example will demonstrate how the Spearman-Brown formula operates. The generalized formula is:

$$r_n = \frac{nr}{1 + (n - 1)r}$$

in which r is the coefficient of reliability obtained between the parts of the divided test; r_n is the reliability of the test n times as long as half the original test.

In the method of odd-even reliability, n is 2, since the original test

has been divided in two equal parts. Assuming, then that the odd-even coefficient (r) is .80, and substituting the values in the formula, the reliability of the whole test (r_n) is found to be .89. This estimated reliability coefficient for the test as a whole is the one usually reported in psychological research and in test manuals.

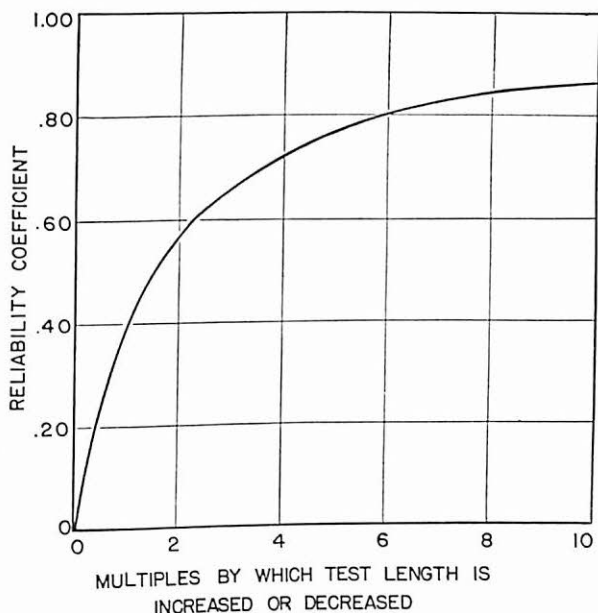


FIG. 1.1. *Changes in Reliability with Changes in Length of Test, as Predicted by Spearman-Brown Formula. Unit Length Reliability is .40.* From L. L. Thurstone, *The Reliability and Validity of Tests*. Ann Arbor, Mich.; Edwards Bros., 1935.

The Spearman-Brown formula may be used to estimate the effect upon reliability of a test of a given length if it should be increased by any multiple (say, 3 or 4 times) or decreased by any fraction (say, $\frac{1}{2}$ or $\frac{1}{3}$). There is a point of diminishing returns, so to speak, beyond which the very small increase in the reliability coefficient, resulting from increase in length, does not warrant the extension of a test. (See Figure 1.1.) Figure 1.2 illustrates increase in test reliability as the length of a test is doubled. This figure demonstrates what happens when reliability is calculated by the split-half method and then corrected by the Spearman-Brown formula.

Selecting odd-numbered items as one half of the test and even-numbered items as the other half is justified on these grounds: items in most tests (as will be seen) are grouped together according to type (number sequences, vocabulary, etc.) and are graduated according to difficulty, from easiest to hardest. Thus, when this systematic arrangement is employed, the odd-even procedure yields very close approxi-

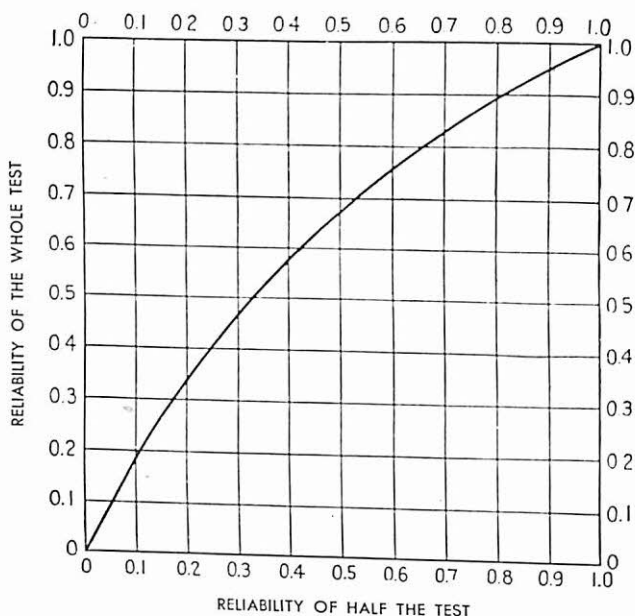


FIG. 1.2. Showing the Increase in Reliability of the Whole-Test Scores as a Function of the Reliability of Half-Test Scores, when the Spearman-Brown Formula Is Applied.

mations to equivalent half-scores, because each half-score is based upon the same types of items and the same number of each type; and each half-score is based upon items which progress in difficulty in approximately the same degree. For example, consider the first ten items of a single type (known as a subtest), say, verbal analogies. Numbers 1, 3, 5, 7, 9 are, as a group, of approximately the same total difficulty as numbers 2, 4, 6, 8, 10—if they are graduated in difficulty from 1 to 10; for both the odd-numbered and the even-numbered include items from practically the entire range of difficulty represented from numbers 1 to 10.

There are other methods of selecting items for getting part scores; but since they are not too frequently encountered, they will not be described here. Whatever method is used to get equivalent half-scores, the procedure must be based upon the psychological rationale and upon the format of the particular test under consideration, in order to insure, as far as possible, equivalence of items in respect to mental processes involved and in respect to difficulty, as well as number of items in each part.

The split-half method tends to overestimate the *predictive* reliability of an instrument, since the correlation is not affected by the ordinary conditions that cause normal fluctuations in a person's performance on different days. In particular, this method should not be used in estimating the reliability of a pure "speed test,"—by which we mean a test whose items are of the same degree of difficulty throughout and which therefore measures only rate of performance at the given level of difficulty. Since all items in the test are of equal difficulty, an examinee should do as well with any one item as with any other. Hence, to measure rates of performance and to differentiate among individuals, the time limit and the length of the test should be such that no one is able to complete all the items. Under the circumstances, except for chance errors in performance, the odd-even correlation should be $+1.00$ (perfect positive), because the test is, presumably, uniform throughout and the psychological function being measured (speed) is operating uniformly on all items. It is apparent, thus, that the total scores on the odd-numbered items should equal those on the even-numbered. One test manual, for example, reports an odd-even reliability coefficient of $.99+$, but the manual also reports a coefficient of $.88$ when the scores of two equivalent forms were used.⁵

The best practice is to use the test-retest method with a highly speeded test. Tests differ in respect to the significance of speed of performance, even when the items are also scaled in difficulty. As the role of speed in a test decreases, the odd-even correlations will differ less and less from those obtained with the test-retest method.

The standard error of measurement. This index is an estimate of the deviation of a set of obtained scores from their "true" scores.⁶ It is

⁵ *Differential Aptitude Tests*, Manual, page C-6, New York: The Psychological Corporation.

⁶ A "true" score is the measure that is quite free from and uncontaminated by chance factors and errors of measurement; theoretically, it represents an individual's true level of performance on the test being used.

dependent upon the standard deviation of the distribution of obtained scores and upon the coefficient of reliability of the test from which the distribution of scores was obtained. The formula for determining the standard error of measurement is written:

$$SE_{(meas.)} = SD_x \sqrt{1 - r_{xx}}$$

in which SD_x is the standard deviation of the distribution of the obtained scores, and r_{xx} is the reliability coefficient of the test.

Assume that the standard deviation of a test (SD_x) is 12 IQ points and that its coefficient of reliability (r_{xx}) is .90. Substituting these values in the formula, we find that the standard error of measurement is approximately ± 3.8 points. This statistic is interpreted as follows: assuming that the test scores are normally distributed and that the "errors of measurement" are similarly distributed, then approximately 68 percent of the obtained scores are *within* ± 3.8 points of the true scores for the persons measured. Otherwise stated, the odds are 68 out of 100 (or 68 to 32) that a particular individual's obtained score is in error by 3.8 points *or less*. Then using the table of probabilities for standard deviation values, we can say, further, that the probabilities are 19 to 1 (95 in 100) that the error of measurement will be 7.6 points (twice the standard error of measurement) or less; and 99 to 1 that it will be 9.5 points ($2\frac{1}{2}$ times the standard error of measurement) or less.

The foregoing technique gives us the means of estimating an individual's "true" score from a set of obtained scores, of which his is one. Using the data of the illustration above, assume an individual's obtained IQ is found to be 100. The probabilities are, then, two to one that his "true" IQ lies between 96.2 and 103.8. (For practical purposes we would say between 96 and 104.) And the probabilities are nineteen to one that it lies between 92.4 and 107.6.

Obviously, the higher the test's reliability coefficient, the smaller will be the error of measurement, and therefore the greater the predictive value of the test. The standard error of measurement provides us, also, with a basis for judging whether or not the scores for two persons represent a true difference or whether they are only deviations from the same, or nearly the same, true scores. For example, if one person gets an IQ of 100 and another person gets one of 96, are these within the range of the same true score or are they significantly different, statistically? Using the data of the illustration given above, we say they are within the range, and they are not sta-

tistically significant. Also, quite aside from any question of statistical significance, a clinical psychologist knows from experience with details of test performance that no *psychological* significance attaches to a difference between IQ's of 100 and 96, or to similar differences elsewhere on the scale.

It is clear that while it is essential to have the reliability coefficient for a test, as an estimate of relative reliability, it is equally essential to have the error of measurement as an estimate of absolute reliability.⁷

Analysis of variance. As already stated, the degree of reliability of a test depends upon the extent to which variations in scores of the testees are attributable to "true" differences among the individuals constituting the group, and the extent of inaccuracies of measurement. A test is unreliable in proportion to the variation of results attributable to factors of test inaccuracy, rather than to "true" differences among the members of the group. The estimate, in the scores of a group, of the proportions of variation due to each of the several factors, is technically known as *analysis of variance*.⁸

In a study of intelligence-test reliability by this method, we would ask: what factors may be important, and to what extent, in producing the obtained differences of scores on two applications of the identical test (or of equivalent forms) to the same group of persons? First: since individuals differ in any population sample, the analysis should estimate the extent to which obtained differences in scores are due to "true" differences in the functions being measured. Second: if there is some general improvement of scores on the second test, it would be necessary to estimate the "practice" effect. Third: since the two foregoing factors would, in all probability, not account for all differences in scores, it is assumed that there are "residual differences" due to errors of measurement attributable directly to the test being used;

⁷ The SE (meas.) is an over-all index, theoretically applicable throughout the range of scores. It sometimes happens, however, that a test measures with less "error" at some parts of the scale than at others. In that case, it is possible to determine at which parts of the scale the "errors of measurement" are larger or smaller. On the Stanford-Binet Scale (1937), for example, the SE (meas.) is 5.2 points for IQ's above 130, but only 2.2 points for IQ's below 70.

⁸ *Variance* is defined as the mean of the squared deviations from the mean score of the group. A measure of deviation is an index of the extent to which individual scores of a group vary from the group's average score. Variance is the statistical term for the square of the standard deviation.

that is, weaknesses or defects within the test. If additional influencing factors could be isolated, their significance in producing the obtained scores would also be determined. Those factors that cannot be isolated and separately analyzed remain as "residual" factors.

Analysis of variance as a method of estimating reliability is preferred by some psychologists, but it has not been widely used.⁹

Reliability is also evaluated at times by means of statistical devices with which may be calculated consistency of performance from item to item within a test.¹⁰ This method introduces the assumption that the test is completely homogeneous as to functions measured; that is, that each item in the test measures precisely the same composite of mental functions as every other item. In most tests this is a doubtful assumption; but if the assumption is warranted, the technique may be used.

Factors affecting the interpretation of reliability coefficients. In addition to the considerations mentioned in connection with the several methods of estimating reliability, there are other factors that must be taken into account in interpreting reliability findings.

Range of ability of the group tested affects the reliability coefficient. If a reliability coefficient is found with a group that has a relatively small variation of the trait or function being measured, the coefficient will be relatively low. If the group has a wider range in the trait or function, the coefficient will be higher. (See Figure 1.3.) Thus, a test having high reliability for a widely varying group does not necessarily have equal reliability for a significantly more homogeneous group of persons. The reasons for this fact are several, one being the nature of the correlation process and the elements in the correlation formula.

For illustrative purposes, suppose that we are dealing with a completely homogeneous group of individuals, with respect to one measure: namely, chronological age. Assume that everyone in the group is exactly ten years of age. If they are an adequately representative

⁹ Since analysis of variance as a method of estimating reliability requires more than knowledge of elementary statistics, it will not be further elaborated here. See R. W. B. Jackson and G. A. Ferguson, *Studies on The Reliability of Tests*, Toronto: Department of Educational Research, University of Toronto, 1941. Also C. Hoyt, "Test Reliability Obtained by Analysis of Variance," *Psychometrika*, Vol. 6, 1941, pp. 153-160.

¹⁰ See G. F. Kuder and M. W. Richardson, "The Theory of Estimation of Test Reliability," *Psychometrika*, Vol. 2, 1937, pp. 151-160.

sample of all ten-year-olds, the range in test score might be from extremely low to extremely high. In this instance, since there is no deviation (or range) whatever in one of the measures (chronological age), the correlation coefficient for the two variables (test score and CA) will be zero.¹¹ Such an extreme instance rarely occurs, but it does demonstrate that when there are possibilities for wide variations in one measure (in this instance, the test score) and very restricted possibilities in the other (in this instance, the CA) the coefficient is lowered.

If the age range were two years instead of one, the coefficient of correlation would still be low, but not zero, because *in general* the members of the older group tend to have higher test scores than do those in the younger. But since there is a wide range of capacity within each group, and overlapping of capacity between the two age groups, the coefficient will be low.

A correlation coefficient reflects the *group* trends of the measures. As persons increase in age, mental capacity increases until maximum development is reached. The correlation coefficient will reflect this fact. But since there are wide differences in capacity within any age group, and since there is considerable overlapping of capacity even among rather widely separated age groups, the coefficient will be affected by these facts also. The result will be a coefficient lower than +1.00 (perfect correlation).

Thus, in correlation estimates of reliability, if the age range is wide,

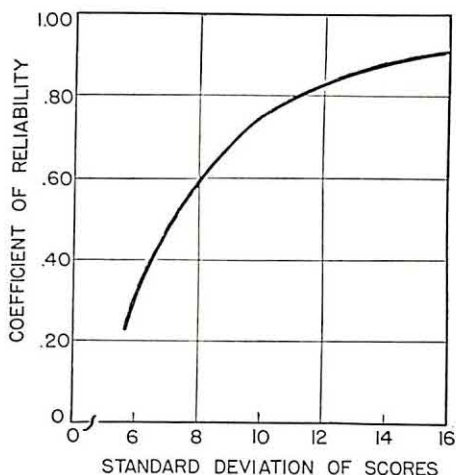


FIG. 1.3. Curve Showing Increase in Test Reliability as Variability of Group Increases.

¹¹ Inspection of the product-moment correlation formula will show this to be the case:

$$r = \frac{\Sigma(xy)}{N(SD_x SD_y)}$$

in which Σxy is the sum of products of the deviations of the paired scores; SD_x and SD_y are the standard deviations of the two sets of measures.



the group trends in scores (higher scores with higher ages) will have increased weight, as compared with a narrower age range in which the age trend has less weight. In interpreting a reliability coefficient of a test, therefore, it is necessary to know the range of ages upon which the test was standardized.

TABLE 1
Raw Scores and Ranks of Students
on Two Forms of an Arithmetic Test

Student	Form X		Form Y	
	Score	Rank	Score	Rank
A	90	1	88	2
B	87	2	89	1
C	83	3	76	5
D	78	4	77	4
E	72	5	80	3
F	70	6	65	7
G	68	7	64	8
H	65	8	67	6
I	60	9	53	10
J	54	10	57	9
K	51	11	49	11
L	47	12	45	14
M	46	13	48	12
N	43	14	47	13
O	39	15	44	15
P	38	16	42	16
Q	32	17	39	17
R	30	18	34	20
S	29	19	37	18
T	25	20	36	19

Just as, in the foregoing illustrations, correlation coefficients were shown to be lowered by homogeneity in one of two variables, so in estimating reliability the coefficient will be lowered by restricting the group's range of variation in the trait being measured. An illustration will help to clarify this matter.¹²

"In Table 1 are shown the raw scores and rankings of twenty students on two forms of an arithmetic test. Looking at the two sets

¹² From *Test Service Bulletin*, The Psychological Corporation, No. 44, May 1952.

of rankings, we see that changes in rank from one form to the other are minor; the ranks shift a little, but not importantly." A coefficient computed from these data is very high: $r = .968$.

"Now, however, let us examine only the rankings of the five top students. Though for these five students the shifts in rank are the same as before, the importance of the shifts is greatly emphasized. Whereas in the larger group Student C's change in rank from third to fifth represented only a 10 percent shift (two places out of twenty), his shift of two places in rank in the smaller top group is a 40 percent change (two places out of five). When the entire twenty represent the group on which we estimate the reliability of the arithmetic test, going from third on form X to fifth on form Y still leaves the student as one of the best in this population. If, on the other hand, reliability is being estimated only on the group consisting of the top five students, going from third to fifth means dropping from the middle to the bottom of this population—a radical change." A coefficient, if computed for just these five cases, is $.50$ (ρ).¹³

"Note that it is not the smaller number of cases which brings about the lower coefficient. It is the narrower range of talent which is responsible. A coefficient based on five cases as widespread as the twenty (e.g., Pupils A, E, J, O, and T, who rank first, fifth, tenth, fifteenth, and twentieth respectively on form X), would be at least as large as the coefficient based on all twenty students." [$\rho = +1.00$]

Furthermore, when the variation among testees is narrow, the correlation between two sets of scores may also be lowered by chance factors and minor psychological factors. Since individuals in such a group are closely clustered—that is, their true differences are small—the changes in scores and relative positions produced by extraneous factors are more significant than they would be in a widely divergent group.

This illustration makes clear the fact that reliability coefficients of a given test may vary as the composition of the tested group changes, even though the performances of the testees themselves are unchanged. Thus reliability data may show that a test discriminates satisfactorily over a *wide* range of the trait or capacity measured; but reliability may still be inadequate where *finer* and *more precise* dis-

¹³ ρ represents the "rank-order" correlation coefficient. It approximates closely the product-moment coefficient (r).

criminations are necessary among individuals who vary within a narrow range.

The practical significance of range and hence of ability is this: in standardizing a test, its author must determine reliability with a group that is similar in average level of ability and in variation of scores to the group with whom the test is to be regularly used. The user of a test should select an instrument that, among other things, provides reliability data based upon a sampling of persons who resemble closely the group of individuals he desires to test and study.

The time interval between testings may be significant in the interpretations of reliability findings. When reliability estimates are based upon odd-even correlations (internal consistency), or upon the scores of two equivalent forms of a test administered at a single sitting or within the same day, the results are uniformly affected by the examinees' physical condition and attitudes, and by the prevailing environmental conditions during the testing. Such uniformity means that the factors external to the test itself are likely to affect both sets of test scores equally, thus increasing the degree of similarity of each person's two scores. This condition tends to give higher reliability coefficients (that is, gives higher estimates) regarding the instrument's predictive value than would be the case if the retest were given after a time interval.

When there is a time interval, the retest results will be affected by the normally expected fluctuations in individual performances and by changes in environmental conditions. Thus, while test results and reliability coefficients obtained at a single sitting or in a single day are most likely to estimate best the consistency of the instrument itself, they do not indicate stability of performance over a period of time as well as do coefficients obtained by the test-retest method, using a time interval. Conversely, the test-retest method is the more likely to underestimate the internal consistency of a test, because factors extraneous to it may affect the scores dissimilarly. The extent to which the accuracy of a test is underestimated by the test-retest method will depend upon the degree to which effective influencing conditions are inconsistent. If the time interval has been quite long, especially in the case of young children—perhaps three months or more—an individual's retest results may be influenced by peculiarities of his growth tempo, or by other more or less enduring conditions such as emotional experiences, which may affect persons of any age.

The longer the interval between tests, the more likely the lowering of the correlation due to intervening factors.

The effects of practice and learning during the interval will depend upon the content of the test being used and upon the examinee's experiences during the interval. For example, if some months have elapsed between two administrations of an educational achievement test, different pupils may have had different amounts and qualities of instruction during the period. The retest scores would, in part, reflect this instructional difference; thus the correlation coefficient would not be solely a reliability coefficient. Or, in the case of a personality test, individuals in therapy or after extensive counseling may have modified their attitudes, values, and behavior sufficiently to produce significant differences in test-retest results.

Which method of estimating reliability is preferable depends upon the problem at hand. Psychologists and educators are usually concerned with knowing: (1) the internal consistency of a test, and (2) the predictive value of a test when it is subject only to the minor or accidental changes in conditions from day to day, rather than to fundamental changes resulting from permanent or semi-permanent changes effected by learning, developmental idiosyncracies, or disturbing emotional experiences. For the first purpose, the odd-even method, or the test-retest method, the tests being given at one sitting or within one day (using equivalent forms), is preferable. For the second purpose, the test-retest method, the tests being given within a week or two (using the same form or equivalent forms), is the preferable one. Under testing conditions that are not too different, the results of the second method will not be far removed from those obtained with the first. A test manual should provide information regarding internal consistency and test-retest results.

Sub-test reliability is not always equal to total test reliability. It has already been explained that, other factors being equal, the reliability of a test increases with increase in length, although not in direct proportion. This principle applies to those scales that consist of several different parts (called *sub-tests*), each of which utilizes a different type of content. Nearly all group tests are of this kind, as are some of the individual scales (e.g., the Wechsler). For these instruments, the total test reliability is higher than that for each of the subtests. It is erroneous, therefore, to assume that the reliability coefficient for the whole may be applied to a part. For example, the Wechsler In-

telligence Scale for Children shows a full scale (nine sub-tests) reliability of .92 for a group of 200 children 7½ years of age, the coefficient having been calculated by means of the split-half technique.¹⁴ Yet the reliabilities of the individual sub-tests, for the same group of children, ranged from a low of .59 to a high of .84. It is obvious that the total score is a more dependable index of the function or functions being measured than is any sub-test score.

Consistency of scorers is a factor in calculating test reliability. Some tests (such as the Stanford-Binet and, in particular, projective techniques) are not entirely objective in scoring, since the examiner at times finds it necessary to judge the correctness or quality of responses. For tests such as these it is necessary to know the extent of agreement in scoring found among two or more competent persons who have scored the same sets of responses. Test authors usually report such data in their manuals; and, in addition, other psychologists will have carried out and reported studies on this problem. Lack of agreement among scorers will adversely affect the reliability findings.

✓ VALIDITY

An index of validity shows the degree to which a test measures what it purports to measure, when compared with accepted criteria. The construction and use of a test imply that the instrument has been evaluated against accepted standards or other criteria which are regarded by experts as the best evidence of the traits or abilities to be measured by the test. Selection of satisfactory validation criteria and demonstration of an appropriate degree of validity is fundamental in psychological and educational testing.

The first necessary condition of a valid test is that it have an adequate degree of reliability. If the reliability coefficient of a test is zero, it cannot correlate with anything. A test that correlates poorly even with itself cannot correlate well with a measure of another variable.

Operational and Functional Validity. It is useful to recognize two kinds of validity, although they are not mutually exclusive. The first is known as *operational* validity; the second is *functional* validity.

By operational validity we simply mean that the tasks required by the test are adequate for the measurement and evaluation of certain

¹⁴ Manual, New York: The Psychological Corporation, p. 13.

specified and defined psychological operations. For example, the Seashore Measures of Musical Talent are actually tests of only *certain essential auditory aspects of musical talent*, but not of "musical talent" which psychologically involves much more than these auditory aspects. Insofar as the Seashore tests differentiate correctly between persons in regard to the specified auditory processes, they are *operationally* valid. On the other hand, these measures are *functionally* valid to the extent that they are efficient in predicting subsequent development of various degrees of skill and competence in the several aspects of music. Thus, the functional validity of a test is the extent to which it is efficient in predicting and differentiating behavior or performance in a specified area under actual working and living conditions.

Numerous other examples can be cited to illustrate the difference between functional and operational validity. Thus, a peg-board test (placing small metal pegs into a perforated board) may well measure manual and digital dexterity (operational), but it might be only slightly useful in predicting "mechanical ability" (functional). Again, a word and number checking test may be quite satisfactory as a measure of perception of details (operational), but it might have limited value in predicting "clerical ability."

It is obvious that functional validity is dependent, at least in part, upon the operational validity of the test. The reason is that the psychological operations required by the test have been included because they have been found to be essential in certain actual situations in which testees will or might be placed. Hence, if the psychological operations themselves are not measured with adequate validity, predictions of later performance will be adversely affected.

Criteria of Functional Validity. The problems of selecting and utilizing satisfactory validating criteria vary with the several kinds of tests. In constructing *tests of general ability* (intelligence), a common practice is to use some or all of the following: scholastic marks, teachers' judgments of individuals' abilities, cumulative scholastic averages over a period of years, number of school grades completed, chronological age, and known groups. The reasons for using these as criteria are: (1) that scholastic records are evidence of mental ability even though influenced by factors other than intellectual ability; (2) that teachers are in a position to evaluate individual ability with some validity, because they observe their pupils over a long period and are able to

make inter-pupil comparisons; (3) that cumulative scholastic averages are more valid than marks or estimates of a single teacher because they represent combined judgments of performance over a longer period of time; (4) that on the whole the more able persons complete more formal education and reach higher levels in school and college; (5) that as individuals grow older their levels of intelligence increase until adult maximum is reached; (6) that definitely known groups, such as gifted, somewhat superior, mentally deficient, slow learning, and average groups will show differential performance on a valid test.

The principal criteria in standardizing *tests of specific aptitudes* (e.g., mechanical, musical) are marks in training courses and differentiation of known groups possessing the aptitude in varying degrees. An example of known groups would be those working efficiently at each of several levels of a mechanical occupation and those in non-mechanical occupations. It is highly desirable, of course, to use degree of success of actual performance in the vocation as an ultimate criterion.

When the criterion of actual performance on the job is used, the following kinds of ratings are obtained: ratings by supervisors, evaluation of the quality of the product, and rate of work. However, the most frequently employed criteria for tests of specific aptitudes are marks and ratings in training courses; that is, criteria of capacity to learn the given skill or the profession, since aptitude tests are used largely to select individuals for training or education in the specified areas, although their use in employee selection is not inconsiderable.

In *personnel work*, in business and industry, where specialized tests are used to select individuals for specific jobs, it is possible, indeed essential to use actual production records or performance ratings as criteria of test validity. If, for example, a personnel department wants to know whether certain measures will identify the potentially best stenographers, the tests might be administered: (1) to a group of employees of several quality levels to estimate the instruments differentiating efficiency; and (2) to newly employed personnel whose performance records, after an adequate period, would be correlated and otherwise analyzed against their test scores.

Tests of educational achievement are validated against school marks and teachers' ratings. Frequently, also, the criterion is "content validity" rather than some external standard. "Content validity"

means simply that the author of the test has determined upon its content by means of an analytical process and his own judgment, as well as that of experts in the subject-matter, as to what is appropriate and germane. For example, in constructing a test of American history, the author examines what he believes to be representative textbooks, consults teachers of history, decides which topics are most significant and what their relative weights should be, and devises items he believes are most representative of these.

Tests of personality traits present an especially difficult problem in validation. Often the author of the test uses "face validity." Whenever an author bases his test upon his own analysis of what is to be evaluated or measured, without reference to prescribed content, as in textbooks, or without subjecting his device to comparison with other external standards, he is using "face validity" as his criterion. At times, the traits presumably being measured by a particular test of personality have been included only by fiat. The sounder tests in this category, however, are validated against actual behavior of the subjects and against clinical diagnoses. But even these criteria present difficulties because they are themselves affected to an appreciable degree by the subjective judgments of the persons making the evaluations of behavior or the clinical diagnoses. These matters will be discussed further in subsequent chapters.

Criteria of validity may be *immediate*, *intermediate*, or *ultimate*. The use of marks in a particular course of study as a criterion in validating a test of specific aptitude is a case of an immediate criterion. The cumulative average marks in an entire training curriculum can be regarded as the intermediate criterion, if performance on the job itself is the ultimate criterion. If a test is being designed solely for the purpose of predicting marks in a single course (say, geometry), then those marks are the ultimate criterion. Whether a criterion belongs in one or another of the categories depends upon the purpose of the test and the number of phases or steps that are available for use as criteria. In fact, when validity findings on a test are to be interpreted, the purposes for which the instrument is designed must be taken into consideration. A particular test may have different validities for different purposes, different age groups, different sex groups, etc. The validity of a test lies in the correctness with which it measures at the time it is administered and in its predictive value for specified activities by specified groups.



Factorial Validity. This method utilizes factor analysis techniques that are not within the scope of the present discussion. Factor analysis theory, however, is discussed in Chapter 3, in connection with theories of intelligence. Yet, since factorial validation is a method used with some tests, students should be familiar with the general nature of the theory.¹⁵

Most tests of mental ability and of personality sample a composite of performances such as verbal knowledge and facility, number facility and quantitative reasoning, memory span, concept formation, etc. Factor analysts maintain that these and others, especially when represented by a single composite index (such as mental age or intelligence quotient) are not "functional unities." Analysts urge that they are not measures of a "pure" ability; that is, just one type of ability uncomplicated by others. Thus, according to this theory, a test is said to have high *factorial validity* if it is a measure of one "functional unity" (e.g., word knowledge) to the exclusion of other elements as far as possible. The factorial process aims to identify, by the method of intercorrelations and further statistical analysis, a list of "functional unities" (also called "primary mental abilities") within a test and the weight contributed by each of these to total performance on the test. The ultimate goal is to devise tests each of which will measure only one "functional unity" and be relatively independent of others (that is, show quite low intercorrelations). Such "pure" tests would then be used singly; or they might be used as subtests in a comprehensive measuring instrument; but even then each subtest is scored and rated independently for the purpose of obtaining a psychological profile for each person.

If validation stops at factorial validity, "operational" validity has been established. But if, after factorial validity is established, we proceed to validate the test against criteria of later performance in working situations, we are making a "functional" validation. The principal contribution of factorial validation is this: instead of validating the total, undifferentiated instrument against functional criteria, an effort is made to identify the component psychological elements and to establish their relative independence, and finally, to correlate these elements separately against functional criteria.

Such analysis into psychological "unities," or elements, is of value

¹⁵ See J. P. Guilford, "Factor Analysis in a Test-Development Program," *Psychological Review*, Vol. 55, 1948, pp. 79-94.

when individuals are to be selected for specialized work or study and their performance predicted therein. For example, since "mechanical aptitude" is not a simple, unitary skill, it is valuable to be able to identify which psychological elements have most predictive value for a specified type of work. Mechanical work may involve a high degree of spatial perception in one situation but not in another; or manual precision and speed; or comprehension of mechanical principles. Also, higher than average intelligence is desirable for the practice, let us say, of both law and engineering. In the former, word knowledge and verbal concept formation are the more significant, whereas in the latter spatial perception and quantitative reasoning are more significant than word knowledge. Factorial analysis can assist in identifying the more limited and immediately relevant aspects of ability required in a given occupation or activity.

Face Validity. This is a term that is used to characterize test materials which *appear* to measure that which the test author desires to measure. Use of the term "validity" in such instances is hardly warranted, for the materials have not been objectively analyzed for validity. In instances of face validity the author and those using the test simply assert in effect that the content of the test appears to be appropriate and to serve their purposes. Face validity is found most often in personality inventories and in some of the more recently published projective tests, notably the Szondi test (described in Chapter 19). This type of "validity" is also found at times in methods used in the selection of industrial personnel. It is, however, unwarranted and should not be resorted to unless a relatively objective approach is impossible or not feasible.

Cross Validation. This term refers to the process of validating a test by using a population sample other than the one on which the instrument was standardized. The reason for using this method is that at times the original validity data may be spuriously high due to the operations of some chance factors that produce a higher correlation than is warranted. As a matter of fact, however, once a test is put to use in a variety of situations and by many different persons, it is being constantly cross-validated; and if it does not prove to have high enough functional value, its use will be, or should be, discontinued.

Methods of Calculating Validity. The most frequently used technique of estimating validity is the *simple correlation* of test scores with each

criterion. A coefficient of a given magnitude cannot be arbitrarily specified as signifying or as not signifying validity. Whenever a validity coefficient is positive and significant, it has some value. In some instances, coefficients of only $+0.30$ have been found useful. Most coefficients, however, should be larger.

ERROR SCORES ON STORE PERSONNEL TEST, FORM FS

TRAINER RATINGS
(Sum of learning ability, working speed, and over-all fitness)

	56 54	53 51	50 48	47 45	44 42	41 39	38 36	35 33	32 30	29 27	26 24	23 21	20 18	17 15	14 12	11 9	8 6	5 3	2 0
12												2	2	1		1	1		
11										1	2		3	3	1	5	1		
10											3	1	3	1	1	2	2		
9									1		5	4	3	6	4	3	1		
8										1	4	4	2	2	2	2	1		
7										3	4	3	1	2	2	4			
6							2			2	6	2	1		2	1			
5								1	1	4	3		1	2	1	1			
4							1	1	3	1	1	1							
3		1				1	1	1			3		1	3					
2								1	1	1		1				1			
1													1						

FIG. 1.4. Chart for Pearson product-moment correlation between number of errors made by 155 grocery store trainees on Part II of the experimental Store Personnel Test, Form FS, and ratings made by the training staff. $r = .46$.

Figure 1.4 illustrates the simple correlation method.¹⁶ The test scores, shown horizontally, were correlated with trainer ratings, shown vertically. The number in each cell shows how many persons earned the scores of that cell, as indicated on both axes. For example, two persons who made between 21 and 23 errors on the test were given trainer ratings of 12; then going to the bottom of the same column, we find that one person who also made between 21 and 23 errors had a trainer rating of 2. For this sampling of examinees, the

¹⁶ From *Test Service Bulletin*, No. 37, 1949, New York: The Psychological Corporation. The data of biserial and tetrachoric correlations that follow are also from this bulletin.

coefficient is .46, which is well within the range of validity coefficients most often found for a single criterion.

The *biserial correlation coefficient* is used when one of the criteria is rated in terms of only two categories: e.g., "pass" or "fail"; "satisfactory" or "unsatisfactory." The second measure, however, is given in terms of variable scores. This method is used when the situation

TABLE 2

Biserial correlation between scores of 52 employed stenographers on the Seashore-Bennett Stenographic Proficiency Test and their supervisors' ratings on stenographic ability. $r_{bis} = .60$. (The Psychological Corporation)

Ratings on Stenographic Ability

Test Scores	Below Average	Average	Above Average	Excellent
19		1	3	2
18		3	3	5
17		1	2	2
16		1	—	2
15		8	5	
14		2	—	
13		3	—	
12		1	1	
11		—	—	
10	2	2	1	
9	—	—		
8	1	1		
Subtotals	3	23	15	11
Totals		26 (Group 1)	26 (Group 2)	

requires only a rough evaluation, as in the illustration presented in Table 2. Here we see that the four groupings on the basis of supervisors' ratings (below average, average, above average, excellent) have been reclassified into two categories (Groups 1 and 2) which have been correlated with stenographic proficiency test scores. The biserial coefficient of .60 indicates that the proficiency test has considerable value in identifying stenographers who will function at satisfactory or highly satisfactory levels.

The *tetrachoric coefficient of correlation* is an index that is found when a coarse classification of two measures is adequate for the pur-

pose at hand. When this index is used, the ratings in *each* measure are grouped into only two classes, providing a "four-fold" table. The data in Table 2 have been so reclassified in Table 3, yielding a tetrachoric coefficient of $+ .60$.

Whether one uses the finer classifications necessary for calculating the product-moment coefficient (the "simple" correlation above) or uses the coarse groupings shown in biserial and tetrachoric calculations will depend upon the nature of the data available and upon the purpose for which validation is to be used.

TABLE 3

Four-fold table for computation of tetrachoric correlation coefficient. $r_{tet} = .60$. (The Psychological Corporation)

		Ratings		
		3-5	6-8	
Test Scores	16-19	6 (11.5%)	19 (36.5%)	High on Test
	8-15	17 (32.7%)	10 (19.3%)	Low on Test
		Rated Low	Rated High	

Multiple correlation is a method whereby two or more criteria are statistically combined and correlated with the test score to yield a single coefficient. Whereas the simple product-moment coefficient indicates the degree of relationship (or co-variation) between two sets of measures, the multiple correlation coefficient shows the relationship between one set of measures (in this instance, test scores) and the composite of two or more other sets of measures (in this instance, the criteria). In other words, while a test might have a low or moderate correlation with a single criterion, it can have a quite significant correlation with several criteria taken together as a composite. This is so because the several criteria in combination have more elements or factors in common with the test than does any one factor taken singly.

Expectancy tables provide a relatively simple, straightforward, and very valuable method of estimating the predictive efficiency of a test.

The estimates are based upon the calculated probabilities that an individual who has a given test score will achieve a specified score or rating in the performance being predicted. We might ask, as examples, the following questions: What are the probabilities that a prospective college student scoring in the highest decile group on a "scholastic aptitude" (intelligence) test will remain in college a given number of terms? What are the probabilities that a child with an IQ of 80-85 will be able successfully to complete the work of the eighth grade? What are the probabilities that a candidate getting an average score on a stenographic proficiency test will achieve a rating of "excellent" or "above average" on the job? Appropriate expectancy tables are intended to answer these and similar questions.

TABLE 4
*Decile Rank on a Scholastic Aptitude Test
and Semesters Completed
(in percents)*

Decile Rank	Terms *						
	2	3	4	5	6	7	8
X	98	95	94	90	89	88	88
VII	94	87	85	82	81	78	78
VI	92	82	79	74	74	73	73
II	85	71	66	61	59	57	56
I	81	67	60	52	51	49	48

* Decile rank X is the highest; I is the lowest

Table 4 is an illustration in point. It presents part of a larger table representing all ten decile groups.

To take two items from Table 4, we may say the probability is that 88 in 100 of the students in the highest decile group on the scholastic aptitude test will complete their academic course, whereas only 48 in 100 of the lowest decile group will do so.

Table 5 illustrates the use of expectancy data in personnel selection. Inspection of this table shows that it may be used to indicate what percentage of individuals obtaining each of the several ratings on actually demonstrated stenographic ability may be found at each of the several levels on the proficiency test. It is also possible to calculate the percentages by rows (instead of columns) so as to indicate the converse: namely, the frequencies of the several ability ratings within each of the score intervals of the proficiency test.

Comparison of Tables 4 and 5 demonstrates that expectancy tables need not be uniform. The form and arrangement of data will depend upon the particular probabilities one desires to determine. But all expectancy tables for tests have this in common: they provide estimates of the probabilities that a certain level or quality of performance may be expected if the test score is known—that is, its functional validity in terms of probabilities in place of or, more often, in addition to a correlation coefficient.

TABLE 5

Expectancy table showing the number and percent of stenographers of various rated abilities who came from specified score groups on the S-B Stenographic Proficiency Test. ($N = 52$, mean score = 15.4, S.D. = 2.9, $r = .61$; score is average per letter for five letters.) (The Psychological Corporation)

Number in each score group receiving each rating on stenographic ability				Stenographic Proficiency Test Scores	Percent in each score group receiving each rating on stenographic ability			
Below Average	Average	Above Average	Excellent		Below Average	Average	Above Average	Excellent
	4	6	7	18-19		17	40	64
	2	2	4	16-17		9	13	36
	10	5		14-15		44	33	
	4	1		12-13		17	7	
2	2	1		10-11	67	9	7	
1	1			8-9	33	4		
3	23	15	11		100	100	100	100

A *cut-off score* is a special instance of the expectancy method. It is a test score that is used as a point of demarcation between examinees who will be accepted and those who will be rejected. For example, Table 6 shows several values from the Cornell Index (a personality inventory discussed in Chapter 17) that might be taken as cut-off scores.

A low score on this inventory signifies fewer personality problems; hence it is more desirable. The table reads: If a cut-off score of 7 on the Cornell Index were used with this group of 1000 persons, 86 percent of those who were rejected after the interview would have been rejected also by the Index; but 28 percent of those accepted after the interview also would have been *rejected* by the index. The other per-

centages are read in the same way. Since the higher scores in this instance are the less desirable, and if 7 is taken as the cut-off level, we mean that a score of 7 or lower would be necessary for acceptance. If 23 were the cut-off, then anyone with that score or lower would be acceptable. In this instance the cut-off score becomes less selective as it increases.

TABLE 6

*Percent of Psychiatric Rejects * and Accepts *
Identified by the Cornell Index (Reprinted
by permission from the Manual. The
Psychological Corporation.)*

Cut-off Level	400 Rejects	600 Accepts
7	86%	28%
13	74	13
23	50	4

* Based upon psychiatric interviews

If we are using a test on which larger scores signify higher and more desirable levels of the ability or trait being evaluated, then the cut-off score becomes the more selective as it is increased. Table 7 is a case in point.

TABLE 7

*Percent of Superior and Inferior Workers Identified by
a Proficiency Test*

Cut-off Score	Superior Workers		Inferior Workers	
	Accepted	Rejected	Accepted	Rejected
20	100%	0%	80%	20%
25	90	10	60	40
30	80	20	40	60

The hypothetical example shown in Table 7 is interpreted thus: If 20 were set as the minimum acceptable score, then all examinees who proved to be superior workers would have been employed; but so would 80 percent of those who proved to be inferior workers. The other cut-off scores are similarly interpreted.

It is clear that cut-off scores are especially useful in instances where many more candidates are available than there are places to be filled, so that the cut-off level may be made highly selective, and where one is not concerned primarily with the individual candidates as such, but

rather with the places, jobs, or niches to be filled. The purpose in using cut-off levels is to identify a maximum number of potentially superior or desirable persons and, at the same time, to eliminate a maximum number of inferior or undesirable individuals. Since no test has perfect validity, and since the true potential of some persons may not be revealed by a single test, screening by means of cut-offs will not be perfect. Some desirable persons will be rejected, whereas some undesirables will be selected. Yet, cut-offs and other methods have a very considerable advantage over subjective procedures previously employed; for they provide the data for estimating with greatly increased objectivity and accuracy what are the chances of identifying the persons with the desired abilities or traits.

Other methods, it will become apparent in later chapters, are used in addition to those already explained. Among these others are, for example, the percent who are successful in adjacent age groups and in groups of known ability; significant increases in scores from age to age and from group to group; closeness of approximation of the distribution of scores to the normal frequency curve. Also, in validating personality scales, extent of agreement by specialists in the interpretation of results is an accepted criterion.

There are instances, too, when very low correlations are regarded as evidence of a test's validity. For example, if one starts with the hypothesis that "mechanical aptitude" is a special ability and, as such, relatively independent of what is measured as "general intelligence," then in constructing a test of the former, one should, among other things, aim to devise a test which has a quite low or negligible correlation with the latter.

Item Analysis. With very few exceptions, psychological tests (other than projective techniques) are made up of a large number of items. The score on each item is added to the scores of the other items to obtain a subtest score or a total score, either or both of which are used in calculating reliability and validity. Ultimately, however, the quality and merit of a test depend upon the individual items of which it is composed. It is therefore necessary, in best practice, to analyze each item in the standardization process in order to retain only those that suit the purposes and rationale of the device being constructed. Item analysis is thus an integral part of both the reliability and the validity of a test.

In evaluating items, three aspects are, in the main, considered: (1) the level of difficulty of each; (2) correlation of each item with the score of a subtest or with the total score of the whole test; (3) the degree to which each item differentiates between a high group and a low group (variously selected), or between several groups at different levels.

The first of these aspects, *item difficulty*, is a matter of the percentage of individuals able to pass each item. In practice, if an item is to distinguish between individuals, it should not be so easy that all persons can pass it; nor should it be so difficult that none are able to pass it.¹⁷ It can be demonstrated statistically that an item passed by 50 percent of a group discriminates between more pairs of persons than does an item passed by a smaller or larger group. For example, if an item is attempted by 100 individuals and passed by only 10, and if the testees are taken by pairs, there are 900 (10×90) combinations in which that item can discriminate between paired members of that group. If the item is passed by 50 in the group, then the number of possible discriminations between paired individuals is 2500 (50×50), this being the largest number possible, as the multiplication of any other proportions will show.¹⁸ Obviously not all items in a scale are or should be such as to be passed by 50 percent of the group. Some are included that are passed by a large percentage and some by a small percentage, with many degrees between the extremes.

There is no formula for determining the exact distribution of item difficulties. A common practice is to select some items whose difficulty is at or close to the 50 percent level, and other items with a wide range of degree of difficulty, in terms of percent passing. If all items selected for inclusion in a test were at the 50 percent level of difficulty, the

¹⁷ Theoretically, it would be desirable that the test be so scaled that there is at least one item which can be passed by all for whom the test is intended. For zero scores on a particular test do not necessarily mean absolute zero capacity in the function being measured; nor will all zero scores necessarily signify the same status. Conversely, it would be desirable that a test be scaled upward to a level where no one for whom the test is intended is able to pass the highest item. This aspect would require, of course, that the test be constructed by a person superior to any of the intended testees.

¹⁸ It is not to be assumed that "50 percent passing" is necessarily the best criterion in placing an item in an age scale (like the Binet), as will be seen later.

Percentages passing an item may be converted into scale values on the base line of the "normal" curve; that is, into standard scores. The assumption here is that the trait being tested by each item is distributed "normally" (bell-shaped curve) in the population being tested.

test would, theoretically, simply divide the testees into two groups: namely, those above this predetermined dividing point and those below it. Such items would not differentiate among the individuals in the group above the 50 percent level, nor among those below it. Hence, for maximum differentiating efficiency, a test must contain items at various levels of difficulty as represented by percentages passing. The final consideration will be the inclusion of items of such a range of difficulty as to yield the highest predictive value when compared with the criteria, taking into account the levels of the ability or trait to be measured and the degree of differentiation to be achieved.

A second method of analyzing validity of individual items is to correlate each item against the score of the subtest of which it is a part (e.g., information, arithmetical problems) to determine whether or not performance on it is consistent with performance on the subtest as a unit.¹⁹ This assumes, of course, that all items in the subtest are expected to be homogeneous; that is, that they measure the same psychological process or combination of processes.

Each item may be correlated, also, against the score of the *whole* test. In that case, the assumption is that all the items throughout the entire test are expected to be homogeneous in basic functions measured. When an item is correlated against the subtest score, it is not necessarily expected to show a significant correlation with the whole-test score, because it may be the intention of the test's author to construct a scale whose subtest scores are relatively independent.

The third technique is to analyze each item in respect to the performance of a low group and a high group; that is, low and high based upon scores on the test as a whole, or upon some external criteria whereby individuals can be classified. As already stated, a very few items should be within the ability range of all, or nearly all, testees. Others should be of increasing selectivity. Some items should, of course, discriminate between two extreme groups, say, the highest and lowest 10 percent of the population tested; but it is desirable to have items whose selectivity extends beyond these narrow boundaries, items that would also dependably distinguish between, for example,

¹⁹ The statistical method used for this purpose is the biserial correlation or the point biserial. See any standard textbook on statistics. For presentation of the problems and methods in item analyses, see F. B. Davis, *Item-Analysis Data*, Harvard Education Papers, No. 2, Graduate School of Education, Harvard University, 1946. Also, J. A. Long and P. Sandiford, *The Validation of Test Items*, University of Toronto Press, 1935.

the highest one-fourth and the second-highest one-fourth, between the lowest one-fourth and the next-lowest one-fourth. Kelley has offered evidence indicating that most marked and significant discrimination between extreme groups is obtained when item analysis is based upon the highest 27 percent and the lowest 27 percent of the group. (The ratio of the obtained difference to the standard error is a maximum.²⁰) This method, however, provides only a crude item differentiation, since it does not provide a basis for differentiating among the large middle group of the population, namely, about 50 percent.

Using this method, one procedure would be to find what percentage of the highest 27 percent and what percentage of the lowest 27 percent passed each item; then, by statistical calculation, to determine if the difference between the two percentages is significant. The same method can be followed with other proportions as well. In fact, items may be analyzed with regard to a wide variety of group classifications. Each item might, for example, be analyzed with reference to high, average, low average, and low groups, classification being based upon total test scores or upon external validating criteria.

Validating Objectives. The objective of all validating procedures is to make the most useful selection of test types and test items from among those available, such as to yield the highest prediction of the criterion or criteria. *The first step in preparing such test items is insight into the psychological processes involved.* The next prerequisite is that the items shall be well and precisely written. Then, basically the ultimate decision as to what are the criteria of validity in any area of testing rests upon the analytical judgments of and agreement among those specialists best qualified to evaluate objectives and behavior, who take into account the purposes for which and the groups for whom the instrument is intended.

²⁰ T. L. Kelley, "The Selection of Upper and Lower Groups for the Validation of Items," *Journal of Educational Psychology*, Vol. 30, 1939, pp. 17-24. For a detailed technical treatment of measurement techniques, especially as related to personnel problems, see R. L. Thorndike, *Personnel Selection: Test and Measurement Techniques*, New York: Wiley, 1949.

2.

INTERPRETATION OF TEST SCORES: QUANTITATIVE AND QUALITATIVE

AN INDEX OF RELATIVE RANK

The raw score (that is, the actual number of units or points) obtained by an individual on a test does not in itself have much, if any, significance. One test may yield a maximum score of 150, another 200, and a third 300. Obviously, then, any point score on one of these tests is not directly comparable with the same number of points on either of the others; a score of 43 on one test cannot be directly compared with a score of 43 on another. Furthermore, the average scores of each of these will in all probability be different, as will the degree of variation of scores (called the *deviation*) both above and below the average. For example, the average (*mean*) score of the first test for a given age is, let us say, 90, with approximately the middle two thirds of the scores falling between 75 and 105. For the second test the mean is, say, 120, with the middle two thirds of the scores between 100 and 140; while for the third test the mean is 180, with the range of the middle two thirds between 150 and 210.

It is clear that if scores obtained on each of several tests are to be compared, indexes must be used which will express the relative significance of any given score; or what is known as *relative rank*. In the example given above, assuming that all three tests are intended for the same group, the mean scores of 90, 120, and 180 have the same relative significance—that is, persons making these scores would be at the average in each. Similarly, scores of 75, 100, and 150 have the same

relative significance in their respective tests; for persons getting these scores would be one standard deviation below the means (averages), which signifies that their scores surpass only about 16 percent of all the scores made by the population samplings upon whom the tests were standardized.

Innumerable other comparable points and scores could be selected for illustration. Obviously, however, such score-for-score comparisons would be extremely cumbersome and would, in each instance, have to be interpreted in terms of some common, meaningful index. Hence, to facilitate interpretation, psychological tests (with few exceptions) provide tables of *age norms*, or *grade norms*, or *percentile ranks*, or *decile ranks*, or *standard scores*. These indexes are defined in the following paragraphs.

Norms. A *norm* is the average or typical score on a particular test made by a specified population—for example, the average (mean) intelligence test score for ten-year-olds, or twelve-year-olds; the mean score for fifth-grade pupils on a test of arithmetic fundamentals. Reference to a table of norms enables one to rank an individual's performance relative to his own or other age groups or grade groups. Thus, for example, a child of ten may attain an intelligence test score that is average for his own age group, or for a population of nine-year-olds, or for those ten and a half, etc.; or a fourth-grade pupil, on a test of arithmetic fundamentals, may score at the level typical for his grade, or for some grade above or below.

Mental Age. By means of tables of norms, it is possible to assign an individual an "age" rating, on the basis of his performance on the particular test being used. Thus, an individual, regardless of his age, who gets an intelligence test score that is equal to the norm of the ten-year-old population would have a "mental age" of ten as determined by that test. If his score equaled the norm of the eleven-year-old population, his "mental age" would be eleven. Hence we define *mental age* (MA) as the level of a person's mental ability as expressed in terms of the chronological age of average persons having the same level of mental ability.

Educational Age. If a pupil obtains a total score on a battery of achievement tests (covering several school subjects) equal to the norm of pupils who are twelve years old, he is said to have an "educa-

tional age" (EA) of twelve. *Educational age* is defined as the age equivalent of an individual's score on an achievement test as shown by age norms for the test in question, measuring achievement in a group of school subjects.

In a similar manner, age levels can be determined for individuals on any test whose purpose is to rank persons on the basis of performance according to age. There are times, however, when it is desirable to know the relative level at which a person is located, in respect only to a more narrowly specified group. The group, for example, might be of a particular age or grade, adults at large, or a particular class of persons such as college freshmen. For this purpose, the most commonly used indexes are percentile rank, decile rank, and standard scores.

• **Percentile Rank.**¹ An individual's percentile rank on a test designates the percentage of cases or scores lying below it. Thus a person having a percentile rank of 20 (P_{20}) is situated above 20 percent of the group of which he is a member; or, otherwise stated, 20 percent of the group fall below this person's rank. A percentile rank of 70 (P_{70}) means that 70 percent fall below—and so on for any percentile rank in the scale. In effect, this statistical device makes it possible to determine at which one-hundredth part of the distribution of scores or cases any particular individual is located. By this means a person's relative status, or position in the hierarchy, can be established with respect to the traits or functions being tested. And, as will be seen, psychological measurement, unlike physical measurement, derives the greatest part of its significance from relative ranks ascribed to individuals rather than from units of measurement.

• **Decile Rank.** The decile rank is the same in principle as the percentile rank; but instead of designating the one-hundredth part of a distribution, it designates the one-tenth part in which any tested person is placed by his score. The term *decile* technically means a dividing point. By "decile rank" we signify a range of scores between two dividing points. Thus a testee who has a decile rank of 10 (D_{10}) is located in the highest 10 percent of the group; one whose decile rank is 9 (D_9) is in the second-highest 10 percent; one whose decile rank is 1 (D_1) is in the lowest 10 percent of the group.

¹ Also called "centile."

Standard Score. This index (Z) is somewhat less obvious in its meaning than percentile and decile ranks, although it, too, designates the individual's position with respect to the total range and distribution of scores. The standard score indicates how far, in terms of standard deviation, a particular score is removed from the mean of the distribution. The mean is taken as the zero point, and standard scores are given as plus or minus. If the distributions of scores of two or more tests are approximately normal ("bell-shaped"), standard scores derived from one distribution may be compared with those derived from the others.

The formula is:

$$Z = \frac{X - M}{SD}$$

in which X is an individual score, M is the mean of the distribution and SD its standard deviation.

Assume, for example, that the mean IQ of a group is 100 and that the standard deviation is 14. In this distribution an individual reaching an IQ of 114 has a Z score of + 1.0. Another individual having an IQ of 79 has a Z score of - 1.5.

TABLE 8

Proportions of Cases, or Area Under the Curve,
Corresponding to Given Standard Scores

Z Score	Approximate Percent of Cases from the Mean
.25	10
.50	19
.75	27
1.00	34
1.25	39
1.75	46
2.00	48
3.00	49.8

Standard scores must ultimately be given percentile values to express their full significance. Since the number of cases encompassed within a given number of standard deviations in a normal distribution is mathematically fixed, it is always possible to translate a Z score into a percentile value. Thus a person having a Z score of + 1.0 has a percentile rank of approximately 84; that is, his score surpasses 84

percent of the scores in the group. The person having a Z score of -1.5 has a percentile rank of approximately 7, surpassing only 7 percent of the scores. Table 8 below shows several standard scores and their percentile values, for illustrative purposes.

As an index of relative rank, the standard score is preferred by some psychologists because it is a well-defined property of the normal curve, representing a fixed and uniform number of units throughout the scale. Percentiles and deciles, on the other hand, are positions of rank in a group and do not represent equal units of individual differences.

Quotients. The use of tables of norms for the determination of the several kinds of test "ages" has already been mentioned.² Now, in addition to these performance "ages," it is a rather common practice to determine a "quotient." Of these, the most widely known is the *intelligence quotient* (IQ), found by the simple formula:

$$IQ = \frac{MA}{CA} (100)$$

in which MA is the individual's mental age and CA is his chronological age. Thus, it is clear that the IQ is the ratio of one's mental age to his life or chronological age (multiplied by 100 to remove the decimal) and indicates *rate of mental development* or *degree of brightness*. If mental development keeps pace with life age,³ the quotient is 100; if mental development lags or is accelerated, the quotient will be less than or greater than 100, depending upon the degree of retardation or acceleration.

Tests of educational achievement, as already stated, yield educational ages (EA). This index may be divided by CA to give an *educational quotient* (EQ); that is, an index showing, presumably, whether a person's knowledge and understanding of a group of school subjects are commensurate with his life age, or whether above or below what would be expected of him for his age.

Educational age may be divided also by MA instead of CA. If that is done, the index is the *accomplishment or achievement quotient* (AQ). The reason for using the mental age instead of the chronologi-

² There are a few tests for which mental ages are not derived from tables of norms, notable among these being the Stanford-Binet Scale and the Merrill-Palmer Scale. These tests and their scoring technique are presented in subsequent chapters.

³ That is, until maximum capacity is reached.

cal age in the denominator is that the former is regarded as the more valid index of learning capacity. Hence, dividing the EA by the MA yields a quotient which shows whether or not the individual is working up to mental capacity as revealed by the intelligence test.⁴

Subsequently, more will have to be said concerning these and other quotients. For the present, however, it is to be noted that quotients, like the other rating devices already presented, are in fact indexes whose significance is to be found to a considerable extent in the relative status they give individuals.

Let us assume that we are dealing with three boys, all of the same age. Suppose that their intelligence quotients are 50, 100, and 150.

Since these are numerical ratios $\left(\frac{MA}{CA} \times 100\right)$, it is natural to assume that they have a *quantitative* significance. So they do—for they indicate rate of mental development. But these quotients also have a *qualitative* significance—for, among other things, they indicate each boy's position in the "hierarchy of intelligence." If the measure of intelligence is valid, the boy having the IQ of 50 is seriously retarded and is in the lowest one percent of the population in respect to the psychological functions being tested; the boy with the IQ of 100 is the "typical" or "average" individual, midway up (or down) in the distribution of intelligence; and the boy having the IQ of 150 is very superior and belongs in the top percentile rank of the group.

Qualitative significance of the intelligence quotient can be illustrated further by asking this question: Is the brightest of these three boys one and one-half times as intelligent as the "average" boy, and three times as intelligent as the retarded one? The fact is that this question cannot be answered in terms of numbers; it is impossible actually to say how many "times" more capable or less capable one is than the others, because *the IQ is not a percent*. But each of these quotients has certain connotations. In this example, the qualified school or clinical psychologist will be able to draw important inferences from each boy's IQ regarding rate and quality of school learning, extent and level of educability, vocational possibilities and levels, and probable types of interests.

The boy with an IQ of 50 probably will not be able to complete

⁴ There are several theoretical problems concerning the AQ (such as the logical fallacy of an AQ greater than 1.00) which will not be dealt with at this stage, but will be treated in Chapter 14.

more than the second grade; the boy having the IQ of 100 should be able to complete twelve grades; the boy with an IQ of 150 will be able to progress in education as far as his interests and motives indicate. Obviously, too, the kinds of occupations that will be open to the first boy are very limited; those open to the second will be numerous; those open to the third will be practically unrestricted, so far as the factor of mental capacity is concerned. And the same may be said of the range of interests in general that will be within the scope of each. These facts are of clinical significance, but at present there are no psychological or statistical means whereby one can calculate how many times more or less capable one person is as compared with another.

A caution is necessary at this point. The inferences drawn in the preceding paragraph cannot be based solely upon the numerical IQ value without reference to the clinical features in the test performances or other factors not shown by the numerical index. We have assumed that there are no complicating factors and that the IQ's are valid measures of the capacities and performances of the three boys. The boy with 150 IQ, however, might be an unstable personality who is failing in most or all of his school subjects. The boy with 100 IQ might have been penalized on the test by linguistic factors. And the boy with an IQ of 50 might show a "scatter" (inconsistency and variation) of performance indicating emotional disturbance rather than intellectual impoverishment. Occasionally, also, it will be found that a high test rating may be attributable to an inconsistently high level of performance on one or a few types of subtests (e.g., memory span, word knowledge), just as, conversely, it occasionally happens that a person's IQ is depressed by an inconsistently low performance on one or a few subtests.⁵

PSYCHOLOGICAL MEASUREMENT CONTRASTED WITH PHYSICAL MEASUREMENT

The indexes thus far presented—ages, percentile rank, decile rank, standard score, and quotients—are intended to emphasize the principle that basically all psychological tests yield results that rank individuals *in relation to* their fellows.

⁵ By "inconsistent" we mean that the individual's levels of performance on these few subtests differ markedly, in one direction or the other, from the general and more uniform levels of his scores on the other subtests.

The raw scores obtained on psychological tests are not comparable with or similar to the values obtained in the measurement of physical traits or phenomena, as, for example in measuring length, weight, or light intensity. In the physical realm, the units of measurement are fixed and constant throughout the entire scale. An inch, a pound, a candle-power—each has the same value and physical significance at whatever place on the scale it is measured. Psychological measurement, by contrast, is more difficult and is confronted by special problems. In the first place, it is not possible to determine the “inherent” difficulty of an item in a psychological test in terms of constant units, as it is possible to find the length or weight of any object. Whereas in the measurement of physical phenomena it can be found that a given object is of X length and Y weight, and hence, let us say, twice as long and three times as heavy as another object with which it is being compared, no such *direct* measurement and comparison are possible in psychological testing. In this realm, the measurement value of a test item is dependent basically upon the percentage of persons able to pass the item in the population group for whom the test is intended.

If in the testing of a particular ability, one item is passed by only ten percent of a group, whereas another item is passed by fifty percent, it cannot be said that the first is five times as difficult as the second, because “percent passing” is not a unit in the sense that an inch or a pound is. What can be said is that an individual able to deal successfully with the first item belongs in the highest decile group, while one who cannot pass the first but is able to pass the second falls at the midpoint or “average” level of the group in respect to that item of the test. This interpretation is significant psychologically and educationally. Or, to use another instance, in the case of the Stanford-Binet Scale the age level at which an item is placed—and hence its value in the scale—is determined by the age group in which the “average” individuals pass that item. Thus, anyone able to solve a reasoning problem placed at the ten-year level, for example, but failing to pass reasoning problems at the eleven-year or higher levels, may be said to have typical ten-year-old ability in respect to that mental task.

This leads directly into the problem of the meaning of “mental age” and other “age” units. *Mental age* may be defined as the level of a person’s mental development expressed in terms of the *chronological age* of average individuals of the same level of mental development.

Or, otherwise stated, mental age is an index showing one's level of mental development, corresponding to the level of mental development of average persons of the coinciding *chronological age*. Thus, if a child's mental age is ten, he has reached the level of mental development attained by average children of ten years, regardless of the actual life age of the child being tested.

Now, suppose there are four individuals having mental ages, respectively, of five, six, twelve, and thirteen. Is the difference between

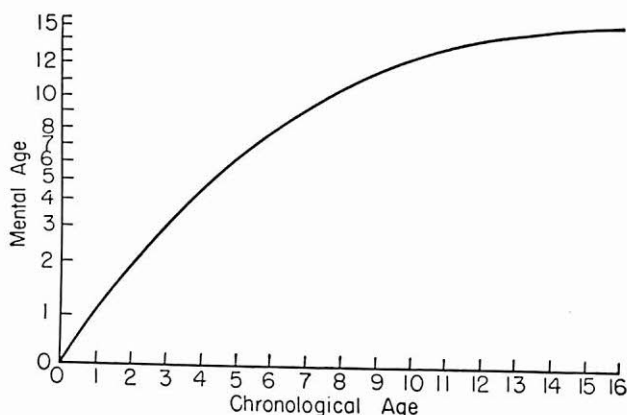


FIG. 2.1. Hypothetical Curve of Mental Growth, Illustrating Decreasing Yearly Increments.

mental ages five and six the same as that between mental ages twelve and thirteen? It is not; for, as measured by mental tests, the rate of mental development at five and six years of age is more rapid than subsequently; hence, the increment between the earlier years is greater. Figure 2.1 is the curve accepted by most psychologists as representing rate of mental development. The outstanding feature of that curve, for the problem under consideration, is the fact that it rises at a *decreasing rate* with increasing age. The curve is "negatively accelerated." Or, psychologically speaking, with each succeeding year, until maximum development is attained, the amount of increment in mental growth and development is less than in the preceding year.

It is obvious, therefore, that each successive year of mental age added to an individual's level represents something less in measurable growth and development than the preceding year's increment. In other words, mental age units are not uniform; they simply rank an indi-

vidual with respect to the average mental capacity of an age group. The same principle—nonuniformity of age units—applies also to all other types of psychological tests which translate their scores into age equivalents.

Although psychological testing would be facilitated and would be given a higher degree of precision if its measuring units were fixed and uniform, the fact is, nevertheless, that the available indexes of *relative* rank are essential educationally and clinically in evaluating the status of an individual's mental development, his educational progress, his particular aptitudes, his social maturity, and even certain non-intellective aspects of personality. Experimentally, too, these same indexes have been indispensable in studying a host of practical and theoretical problems such as sex differences, effects of environmental conditions, inheritance of intellectual capacity and of special aptitudes, occupational differentiation, racial differences, relationships between physical and mental development, and many others.

CLINICAL ASPECTS

Scores, whether raw or converted, do not suffice for the complete interpretation of an individual's performances on psychological tests. The several aspects of test standardization thus far presented are concerned with the performance of *groups* of persons and with *average relationships* revealed by statistical treatment of results. It happens, however, that while certain types of test items meet some or most of the statistical requirements of validity, they are unsatisfactory as indicators of intelligence when used for clinical purposes. For example, on the Stanford-Binet scale, the percentage of adults able to repeat eight digits forward (*digit-span* test) is approximately the same as the percentage who can solve one of the more difficult reasoning problems. Yet, in clinical examinations psychologists find some adult mental defectives who can pass the former test, though it never happens that a true mental defective can succeed with the latter. What this means, in effect, is that statistical validation of a test item is not always sufficient; it must be supplemented by the pragmatic criterion of use with a wide variety of individuals in a variety of situations to show whether or not it has discriminative value as between individuals at the several levels of ability.

Psychological tests, as already noted, are standardized on the basis of the performance of a representative population; and an individual's

rating is determined by the relationship of his performance to that of a group as a whole. Thus we have the several "ages" (e.g., mental age) and "quotients" (e.g., intelligence quotient), percentile and decile ranks, and standard scores. Any useful test should yield one or more of these ratings. In more recent years, however, without denying the usefulness and value of these indexes of relative status, increasing emphasis has been placed upon "patterns" of performance as clinical aids to psychological diagnosis and counseling.

A person's responses to tests are now frequently analyzed for the purpose of discovering whether he shows any special abilities or disabilities, whether there are marked discrepancies between responses on some types of materials as against responses on others, whether certain psychological processes seem to be impaired or are markedly superior to others within the individual. A general contrast, for example, might be found between tests involving verbal materials and those which are nonverbal in character; the associative processes might be disturbed; memory or spatial perception might be found to deviate markedly in one direction or another from an individual's general level of capacity. Recent investigations have indicated that "patterns" of response may be useful in differentiating and diagnosing the several categories of maladjusted and abnormal personalities, as well as for discerning more clearly the mental defectives.

Also, it has been found that persons of equivalent general mental status may have different "patterns" of performance, or abilities, which in sum, nevertheless, give them much the same over-all and general ratings in terms of a single index (e.g., mental age, percentile rank). That is to say, it is possible for two persons to have test ratings which are *numerically* similar but whose "mental organizations" are dissimilar, since the components of each total rating differ to a greater or lesser degree from those of the other.

If, therefore, the psychologist's concern is not primarily with group trends or averages but rather with a particular individual, he will, to be sure, want to know the age level of performance and the consequent "quotient"; but he will also analyze the details of the individual's performance for the purpose of discovering that person's particular pattern or idiom, in order to discern his particular form of mental organization, specific evidences of retardation, or disability, if any, and details of his development.

In more recent years there has been a partial shift in emphasis from

almost exclusive concern with the analysis of abilities and methods of psychological measurement, as such, to an examination of individual performance and individual idiom, and to the individual as a functioning and dynamic unit. After all, any given test measures only a segment of a total personality; that segment is an integral part of the totality and is influenced by the whole. Hence, the psychologist who is concerned with insight into the nature of an individual's abilities must be able to *evaluate* a person's performance as well as *measure* it. The data and indexes derived from psychological tests are, for the most part, objectively determined; but their clinical use involves judgment, subjective assessment and interpretation, based upon a variety of data from several sources. The experienced clinical examiner will supplement the test's numerical results with his observations of the testee's attitudes during the examination and the manner in which he attacks the problems of the test: his degree of confidence or dependence, his cooperativeness or apathy, his negativism or resentment, the richness or paucity of his responses. The individual test situation can thus be, in effect, an occasion for general psychological observations—really a penetrating psychological interview.

Ability not only to score a test but to assess and interpret responses and to evaluate the individual's behavior during the examination is a clinical art that the psychologist develops from working with persons rather than with tests alone, though for the practice of his art he must, of course, thoroughly understand the psychological and statistical foundations and hypotheses upon which the tests are based.

A few specific instances of the *qualitative* analysis and interpretation of test responses will illustrate the kinds of observations that constitute the clinical aspects that supplement numerical scoring.

Word definitions are generally acceptable at a fairly elementary level; but they vary in level and quality from purely concrete, to functional, to conceptual or abstract. Differences in quality level are indicative of differences in modes of thinking. It happens at times, also, that some words are emotionally charged for the examinee, in which case his definition and behavioral response may be revealing.

Some test items permit the exercise of considerable freedom in response. These responses may reveal the examinee's attitudes, values, and modes of meeting life situations. In this category are test items that ask, "What is the thing to do when . . . ?" Or, "Why should we . . . ?" The subject's reactions to such items, the qualities of his verbalizations in making the responses, and the presence or absence

The findings on a test—whether it is of general intelligence, specific aptitude, personality, or school learning—indicate the present status of each person examined. They do not, however, tell the psychological examiner by what course the person arrived there; nor do they indicate specifically what factors were operative in his development. The clinical approach, while accepting and utilizing standardized tests and norms, insists upon viewing and evaluating any given individual's performance and status in the light of a variety of other measurements, observations, and activities of that individual, and upon interpreting the objective quantitative data according to the part they have in the total. For instance, children who have suffered from prolonged and serious nutritional disturbances or deficiencies or who are suffering from severe anemia will appear listless, apathetic, and deficient in mental capacity according to standardized tests. Other children, of apparently retarded mental development, may be suffering from serious deficiencies of vitamin B complex, while still others may measure at a level of retardation because of emotional pressures and "block-ing." ⁷ Furthermore, the performance of some children on standardized tests is of inferior quantity and quality because they developed under conditions of psychological impoverishment, whereas test norms are based on the assumption that all individuals being examined have had approximately equal opportunity, in the grosser aspects of environment, for mental development. Often, of course, that is not the case. These facts, and others of the same kind, indicate that in the case of some individuals, performance and consequent relative status may be impaired by nutritional deficiencies, by emotional handicaps, or by other unfavorable environmental conditions.

The fact that psychologists are tending to put increasing emphasis upon the significance of the individual pattern in test performance (especially in diagnosing cases of behavior and educational maladjustment) and upon the individual as a whole does not mean that statistical and group studies are unnecessary. Such studies are essential in providing norms against which any individual's performance may be projected in the process of evaluation, in giving more precise meaning and significance to any single score, in demonstrating the great range of human variability in any trait or function, and in providing the

⁷ A condition in which the functioning of mental abilities is impeded due to the individual's emotional state or mental conflict.

of strong feelings reveal some of the nonintellective aspects of his personality.

The subject's specific comments while performing a task are of possible significance in regard to his attitude toward himself, or toward an authority figure (the examiner), or toward other individuals and institutions in his environment.

Responses to items or random comments may reveal hostilities and anxieties, or wholesome cooperativeness and security.

The manner of speech; the use of expletives, halting and fumbling, restless movements, blushing; or, on the other hand, ease of speech, a relaxed attitude, mild criticism of one's own performance provide valuable clues to the testee's personality.

Character disorders may be indicated by impetuous and uncritical responses that are incorrect but are given with assurance and pre-tentiousness.

The subject's ability to direct his attention toward, to concentrate upon, and to organize a task are often revealed by his mode of approach to a test problem.

The selective character, if any, of a person's vocabulary and information (two subtests widely used) will shed light upon his experiences, interests, cultural background.

A personality trait such as compulsiveness (as opposed to desirable thoroughness and self-criticism) may be revealed by excessively detailed responses and by numerous and unnecessary alternative responses.

Some types of responses indicate pathological or psychotic states: for example, erroneously bizarre responses by an otherwise intelligent person (e.g., London is in Africa; the population of the United States is 1,500,000), by disjointed and irrelevant responses, and by distorted interpretations of the task or problem.

Detection of organic damage through selected kinds of subtests; e.g., disturbance of the visual-motor function as indicated by the diamond copying test (Stanford-Binet) and the object assembly test (Bellevue), among others.

Scatter analysis (discussed in detail in Chapter 15) is essential to the discernment of superior, inferior, and impaired psychological functions.

Sensitized observations on the part of the examiner will enable him, in general, to evaluate *how* the subject proceeded in both success and failure.⁶

⁶ Illustrations of these qualitative interpretations will be given in Chapter 14, on clinical aspects.

means of more precise study of interrelationships among psychological traits and functions.

DIFFERENCE BETWEEN NORMS AND STANDARDS

Norms, as already explained, are average scores or values determined by actual measurement of a group of persons who are representative of a specified population; for example, all twelve-year-old boys, all fourth-grade children, all native-born male adults. Norms, therefore, are averages *obtained under prevailing conditions*—good, bad, or indifferent. These norms may well “. . . reflect all the sins of omission or commission in their [the people’s] nurture and must be critically examined lest we set up as desirable norms for achievement what are but accidental outcomes of our unsystematic and unenlightened nurture of children . . .”⁸ In other words, a norm of psychological performance or of a physical trait is not necessarily one with which we should be satisfied; for it reflects development under conditions that may be and often are much less than optimal. As an example, consider the “average vocabulary” of the eighth-grade pupil. The average, or norm, of this group will be dependent in part upon their opportunities from earliest childhood for the acquisition and use of language. Their opportunities might have been extremely poor, moderately satisfactory, very good, or at any other level between these three. Norms of performance in respect to some of the psychological processes measured by means of tests of intelligence and of specific aptitude are likewise dependent upon conditions and opportunities present during the course of development. Norms of height, weight, and other body measurements will also reflect past conditions of nutrition and health.

It is necessary, therefore, to distinguish between norms, on the one hand, and standards, on the other; for a standard is the *desired* goal or objective, which may well be above the obtained norm and can be achieved only under improved conditions of development. It is possible that the grade-norms for reading rate and comprehension are below what they could be under improved educational conditions and teaching methods; that age and grade norms for numerical ability are below what they might be; that universal nursery school and kinder-

⁸ L. K. Frank, “Research in Child Psychology: History and Prospect,” in *Child Behavior and Development*, R. G. Barker et al., editors, New York: McGraw-Hill Book Co., 1943, p. 9.

garten experience would promote children's perceptions of form and color and would improve their motor skills, and hence raise the norms; that universal optimal nutrition would raise age norms for height and weight, etc.

Psychological tests measure traits and functions as they exist under present conditions. They do not provide the psychologist and educator with an index of what *ought* to be, except by implication and insofar as obtained results might raise certain suspicions, doubts, and queries in the minds of investigators.

FACTORS IN SELECTING A TEST

By way of summary, the following factors are given as those to be considered in selecting a psychological test.

Norms. The test must provide appropriate and accurate norms, whether they be in the form of age, grade, percentile rank, standard score, or any other type. Norms should be meaningful with regard to the purposes for which the test is intended and to the groups of persons with whom it is to be used.

Administering and Scoring. The procedures of administering the test should be objective and the test items should be amenable to relatively objective and simple scoring, insofar as the nature of the instrument permits. Individually administered tests, like the Stanford-Binet, at times require insightful judgment in scoring responses. Interpretation and evaluation of responses are even more significant in the scoring and analysis of projective techniques for assessing personality.

Time Requirements. The length of the test should not be so great as to produce boredom, satiation, or negativism; for when these set in, the subject does not perform at his best level. Specific time limits cannot be prescribed for all tests or for all types of testees; but in general, shorter time requirements are indicated for younger children and for the mentally retarded. In the case of both of these groups of subjects, the attention span is relatively brief; hence, it may be necessary at times to complete an examination in two sessions.

Interest Level. Test items should be of sufficient interest to motivate the individuals for whom they are intended. Particular items and types of problems devised to measure given functions must be suitable to the

age levels of examinees. Thus, in constructing an intelligence test for the entire range of adult capacity, from very low to very high, it is necessary that the items placed even at the very low levels should be of the sort that will interest an adult rather than a child, even though these low-level adults may be inferior to some children so far as mental capacity is concerned.

The Population Sample. The manual of a test should state in detail the nature of the population sample on which the instrument was standardized and upon which norms are based. The information given should include the following: total number of cases, age range and number at each age level, number of each sex, geographic distribution, socio-economic status and number in each category. For some tests, it will be relevant, indeed necessary, to have information on some of the following: school-grade distribution, number of years of schooling completed, amount and kind of special training (especially for tests of specific aptitudes), special or "abnormal" adjustment problems and history (especially for tests of personality). In short, the prospective user of a test must be certain that the test has been standardized on an appropriate sample of the population and for the same or similar purposes as those contemplated by the prospective user. This principle seems axiomatic; yet it is not always given due consideration.

The Functions or Traits Measured. The test manual should not only state the purpose of the instrument; it should also provide, so far as possible, an analysis (psychological and statistical) of the functions or traits being measured.

Reliabilities. Coefficients of reliability should be provided not only for total scores but for part scores as well, wherever possible. Also, reliability coefficients are desirable for each of the several age-levels and ability levels included within the range of the test. Furthermore, the manual should state which method or methods have been used in calculating the test's reliability. Here, the prospective user of a test must look for information that will help him answer the question: "Reliable for whom and for what purposes?"

* **Validity.** Data on validity are of several kinds, in addition to coefficients of correlation; e.g., expectancy tables, known groups, significance of differences between age levels, etc. The test manual should

explain the characteristics of the criterion *groups*, the nature of other criteria used, the validity of the total test, and the validity of the subtests. It is desirable, also, to have data regarding validity at each of the several age and ability levels. Here again, an answer must be sought to the question: "Valid for whom and for what purposes?"

Reports of Experiments. An ideal to be aspired to is to have test manuals (subsequent to the first or earliest editions) include summaries, findings, and interpretations of the most important experimental studies to which the test has been subjected by various psychologists. Such information will help users to understand more fully the nature of the test and the factors affecting performance on it, thus making for sounder interpretation of results obtained by those who use it. For example: What is the influence of cultural factors? Of practice? Of time limits? Of psychotherapy or counseling?

Psychological tests are scientifically constructed instruments, based upon psychological and statistical principles, as explained in the preceding pages. Familiarity with these principles should provide students with a sounder comprehension of both the values and the limitations of tests than they would obtain from using and interpreting these instruments in a mechanical manner. It is also true that when we test human subjects we are dealing with individuals who do not behave like mechanisms under complete control, with all environmental forces likewise under control and measurable. On the contrary, human behavior is often subtle and the psychological forces motivating or influencing persons in a test situation may be elusive and difficult of evaluation. Furthermore, as has been indicated, since the quantitative data of psychological testing are not as definite, precise, and uniform as are the data of physical measurements, the interpretation of test findings is the more difficult. For these reasons, we have emphasized not only the well-defined scientific principles and procedures of testing, but we have stressed the *qualitative* and *clinical* aspects which are essential, if test findings are to be of the greatest value to the individuals examined.

DEFINITIONS AND ANALYSES OF INTELLIGENCE

DEFINITIONS OF INTELLIGENCE

If intelligence is to be measured and assessed, it is necessary to define it, at least tentatively. A variety of definitions have been given by psychologists; but, as a matter of fact, each of them can be classified into one of several groups.

One group of definitions places the emphasis upon *adjustment or adaptation of the individual to his total environment*, or to limited aspects thereof. According to definitions of this type, intelligence is general mental adaptability to new problems and new situations of life; or, otherwise stated, it is the capacity to reorganize one's behavior patterns so as to act more effectively and more appropriately in novel situations. Thus, the more intelligent person would be one who can more easily and more extensively vary his behavior as changing conditions demand; he has numerous possible responses and is capable of greater creative reorganization of behavior, whereas the less intelligent person has fewer responses and is less creative. The more intelligent person, accordingly, can deal with a greater number and a greater variety of situations than the less intelligent; he is able to encompass a wider field and to expand his area of activity beyond that of the less intelligent.

A second type of definition states that intelligence is the *ability to learn*. According to this definition, then, a person's intelligence is a matter of the extent to which he is educable, in the broadest sense. The

more intelligent the individual is, the more readily and extensively is he able to learn; hence, also, the greater is his possible range of experience and activity.

Still others have defined intelligence as the *ability to carry on abstract thinking*. This means the effective use of concepts and symbols in dealing with situations, especially those presenting a problem to be solved through the use of verbal and numerical symbols. Binet's conception of intelligence belongs largely in this category, for he maintained that it is the capacity to reason well, to judge well, and to be self-critical.

It should be apparent that the three foregoing categories of definitions are not and cannot be mutually exclusive. For the most part, their authors differ in emphasis. Obviously, ability to learn must provide the foundation for adjustment and adaptation to changing or new conditions. And a person may be expected to have learned more or less from situations he had encountered and to which he had made adjustments previously. For if this were not the case, he would have to start anew in every situation which confronted him; there would be no difference between the behavior of an experienced person and that of a novice.

There are, of course, individual differences in respect to learning capacity and in ability to retain, interpret, organize and apply what has been learned; thus previous experiences will have different significance and different learning-value for different persons. And it is learning capacity that constitutes the basis of adjustment and adaptation, although, as will become apparent in later chapters, important nonintellective factors affect adjustment and adaptation.

Yet learning capacity, in the sense only of acquisition of information and knowledge, is not a sufficient criterion by which to evaluate a person's intelligence. Psychologists and laymen alike are agreed that a person who can reorganize and apply what he has acquired for the purpose of dealing with varied and novel situations is more intelligent than one who is capable of little beyond repeating what he had previously acquired, or than one whose behavior follows stereotyped patterns without insight into the essential elements and relations of each new situation. Thus a definition of intelligence as the capacity to behave appropriately and effectively in new situations and a definition of intelligence as the ability to learn are in fact two aspects of the same process.

The third type of definition is also inseparable from the other two. A person *learns* abstractions—principally verbal and numerical—through experience, through contact with and perception of the objects, events, qualities, relationships, etc., for which the symbols stand. Thus, the word “dog” has meaning for a child because it has come to represent a class of objects with which he has become familiar. The word “green” represents a quality he has perceived as an aspect of a variety of objects. The word “charity,” for the individual who has developed sufficiently to understand the concept, has a certain connotation by virtue of the fact that he has experienced events which have been labeled as charitable. The number “five” is meaningful to a person when, as a result of experience with concrete objects, he apprehends the term as representing not only ordinal position but summation as well. Furthermore, if it is to be said that an individual has fully learned to deal with the symbols of abstraction, then it must be true that he understands that the *word is not the thing or the quality* for which it stands. He understands that words and numbers are abstractions which represent objects, events, qualities, relations, etc., but which, in thinking, can be dealt with *as if* they were the things themselves. This aspect of intelligence—the ability to use symbols—is itself the result of an individual’s development and learning. And in its turn, the mastery and utilization of symbols promotes further learning—for it is hardly necessary to labor the point that without language and number, the range of one’s learning would be seriously restricted.

Ability to carry on abstract thinking, it is easy to see, contributes to a person’s ability to adjust or adapt to changing or new situations, because through the use of symbols we are enabled to think through a problem without spending time and effort on sheer trial and error in action; we are enabled to marshal, evaluate, and deal with past experiences; and we are enabled to project our thinking forward. In other words, through the use of symbols and abstract thinking, man is able considerably to enlarge his range of behaving and adjusting, to extend his horizons, and to transcend the immediate concrete and specific situation.

TWO COMPREHENSIVE DEFINITIONS

In more recent years, two definitions of intelligence have appeared which, in effect, combine and extend the three types of defi-

nitions already presented. One writer states:¹ "Intelligence is the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment." The reader can readily compare this definition with those already presented and analyze it with a view to discerning similarities and differences. It will be noted, of course, that this definition encompasses the other three. Although it does not specifically mention learning ability, yet that is surely implied. Two new aspects, however, are added. The definition specifically states that an individual's intelligence is revealed by his behavior as a whole ("global"), and that intelligence involves behavior toward a goal, which may be more or less immediate ("purposefully"). A third aspect is presented by the author in his elaboration of the definition; namely that "drive" and "incentive" enter into intelligent behavior. This aspect is probably included and implied in capacity "to act purposefully" and "to deal effectively" with one's environment, as stated in the definition.

The inclusion of "drive," "incentive," and the like as aspects of intelligence is of very doubtful validity; for to do so is to confuse the issue, the testing instrument, and the results obtained. It is true, of course, that effective utilization of a person's intelligence depends upon the extent and degree to which he employs it. Nevertheless, a single testing device which attempts to combine the measurement of intellectual with nonintellectual traits, without providing for differentiation between the two, would not succeed adequately in either respect.² This is not to say that in assessing an individual's intelligence and personality as a "whole" we should ignore "drive," "incentive," "interest," etc.; for the competent psychological examiner will and must evaluate these and other nonintellectual traits in presenting his test results. Furthermore, as will be seen in later chapters of this book, special psychological instruments are available for the evaluation of nonintellectual traits of personality which the clinician may use to supplement results of intelligence tests if he believes they are necessary.

¹ D. Wechsler. *The Measurement of Adult Intelligence*, Baltimore: Williams and Wilkins, 1944, p. 3. Actually, the usefulness of Wechsler's Bellevue scale is not dependent upon his conception of intelligence. This scale does not, in fact, incorporate all aspects of the definition.

² The Rorschach test attempts to measure both nonintellectual and intellectual traits of personality. This test is discussed in Chapter 19.

Stoddard³ offers the following definition: "Intelligence is the ability to undertake activities that are characterized by (1) difficulty, (2) complexity, (3) abstractness, (4) economy, (5) adaptiveness to a goal, (6) social value, and (7) the emergence of originals, and to maintain such activities under conditions that demand a concentration of energy and a resistance to emotional forces." Here again, the reader will note that this definition does in fact include the first three types of definitions presented; but it goes beyond these in several respects. The author specifies the several attributes of intelligence, and in his enumeration are several not included in earlier definitions.

Degree or level of "difficulty" is implied in all definitions; but Stoddard's contribution here lies in the fact that he rightly insists we must, in testing, distinguish between true differences in degree of difficulty and differences that only seem to exist, as between two or more test items, whereas, in fact, there are no inherent differences in difficulty. For example, the accumulation of pieces of rare information and the ability to define unusual words are not in themselves true measures of difficulty; they may reflect only differences in experience. On the other hand, however, over and above disparity in experiences between various age groups, true differences in difficulty do exist between *problems* that can be solved, let us say, by a group of "average" ten-year-old children, and those which can be solved by an "average" group of eight-year-olds or nine-year-olds.

"Complexity" refers to the number of different kinds and varieties of tasks that can be successfully dealt with. According to this attribute of intelligence, the individual who is able to deal successfully with several different kinds of tasks, at a given level of difficulty, is more intelligent than a person who can successfully undertake fewer kinds of tasks at the same level of difficulty. "Complexity," however, means not simply the addition of one type of performance to others; on the contrary, it means the capacity to assimilate new abilities, to integrate them with others, and thus to reorganize one's patterns or forms of intelligent behavior.

"Abstractness"—that is, operating with symbols, especially at levels of analysis and interpretation—has already been discussed. For Stoddard, this attribute "lies at the heart of intelligence as defined."

"Economy" refers to the rate at which mental tasks are performed

³ G. D. Stoddard, *The Meaning of Intelligence*, New York: Macmillan, 1943, p. 4.

and problems solved. Assuming that the problems are solved equally well, that the solutions are equally effective, the individual working more rapidly would be regarded as the more able, according to this attribute. Acceptance of "economy" as an attribute of intelligence means that tests would impose time limits which should differentiate among individuals in respect to their rates of performance of tasks and solutions of problems at given levels of difficulty and degrees of complexity.

"Adaptiveness to a goal" implies an approach that is more than aimlessly meeting and solving new situations as they arise. This attribute means that intelligent action is directed toward a goal or a purpose. The more comprehensive the goal, the larger and more complete the purpose, the more is intelligent action required.

The student, after examining representative tests of intelligence, might well question whether they do, or even could, satisfactorily test this last attribute; or whether the problems and tasks included in the tests are rather oversimplified and segmental examples of problems and courses of action that a person has to confront and deal with in actual life situations. If the test items are of the latter kind, then their value and validity as measures of intelligence must be shown by the fact that they do indeed predict to an adequate degree the manner and effectiveness with which the testee will deal with and solve actual life situations of broader scope. In other words, what are the predictive values of the items, tasks, and problems included in a test?

The inclusion of "social value" as an attribute of intelligence is of doubtful validity, and debatable at best; for this criterion is essentially moral or ethical, or a matter of subjective evaluation. The basis of "social value" is group acceptability. If this attribute were applied in evaluating intelligence, we should have to minimize our estimates of the intelligence of individuals whose thinking and solutions of problems are not necessarily consistent with accepted social forms, though they might be "ahead of their time"; and of individuals who are capable of difficult and complex mental operations, but whose mental activities lead to no apparent or demonstrable practical and social values. While we may well value more highly the individual whose mental operations culminate in desirable, acceptable, and useful social outcomes, the inclusion of "social value" as an attribute of intelligence would confuse attempts to measure the other, and valid, attributes by injecting largely subjective conceptions of what is socially acceptable

or unacceptable or indifferent. It will be seen later that "social value" is hardly present in current tests of intelligence; although, of course, some psychologists, like Stoddard, take the position that it should be.

"The emergence of originals" as an attribute of intelligence is the ability to create something new and different; it is a characteristic of a high order of thinking and of individuals at the superior end of the distribution of intelligence. Examples of this attribute in operation are the development of a new scientific principle, the discovery of unique relationships in observed data or phenomena, the development of a new machine-design, the development of a new technological process, the new organization and new interpretation of historical or social facts, a creatively original painting or musical composition. It is undoubtedly true that current tests of intelligence provide little opportunity for the measurement of emergence of originals. The question, then, so far as these tests are concerned, is whether this attribute, creative originality, is really dependent upon a combination of abilities which are actually measured by available tests and whether the results obtained by means of the tests are indicative of the degree to which a person possesses this attribute. Some psychologists, Stoddard among them, believe that at present the tests of intelligence do not satisfactorily discern and rate an individual's intellectual originality. Others maintain that the hierarchy of abilities established by means of the tests enables us to identify the persons who possess originality in greater or lesser degree.⁴

Stoddard's last two conditions of intelligent behavior—"concentration of energy" and "resistance to emotional forces"—are subject to the same criticism as Wechsler's inclusion of "drive" and "incentive." Motivation and ability to exert sustained effort are usually regarded as nonintellectual aspects of activity and are certainly recognized as playing highly important roles in one's general effectiveness. But to introduce them into a test of mental ability would be to confuse and probably to invalidate efforts to arrive at a reasonably valid measure of the level of intelligent activity at which a given person is *able to*

⁴ The observation of psychologists is that people who show creative originality, almost without exception, score high or very high on intelligence tests; but many persons may score very high on these tests without having exceptional powers of originality. We need tests of originality, but in view of the very nature of the concept and its expressions, such tests cannot very well be standardized.

operate, regardless of whether he actually does operate at that level in all situations.

While tests of intelligence do not directly measure motivation and concentration of energy on the solution of problems, the psychological examiner does in fact try to develop or encourage conditions wherein the persons being examined will operate at their maximum levels of ability. This can be more nearly achieved in administering an individual test than in administering group tests. Furthermore, if an individual is not adequately motivated, is not expending a maximum of energy during the test, or is handicapped by emotional factors during the testing, these conditions can be discerned much more readily during the examination of one person at a time than during the examination of a group all at once. Indeed, during group testing there may be instances of persons whose test results are vitiated by the effects of the unfavorable conditions mentioned without that fact being known to the examiner. This possibility is a disadvantage of group testing.

Of course, no single test or series ("battery") of tests can provide an unfailing index or a guarantee of motivation, energy output, or freedom from emotional blocking in all future situations requiring intelligent behavior. For man is not a static being; nor does the environment in which he lives remain static. Long-term motives and immediate incentives will change; values and interests will change. The affective (emotional) quality of a person's experiences will influence his subsequent behavior, including situations requiring the utilization of intelligence. Tests of intelligence now in use are not intended to determine the extent to which an individual will in the future concentrate his energy on problems demanding the use of his intelligence, nor to determine whether it is probable that he will be able to remain free from emotional blockings. A variety of personality rating scales and inventories and projective techniques have been devised to evaluate these nonintellectual traits. Although tests of intelligence can and will no doubt be improved so that greater demands will be made upon concentration of attention and sustained effort than is the case with most tests at present, psychologists believe they are warranted in assuming that a qualified examiner will be able to determine whether or not a given person's performance on an *individual* test represents his maximum level at that time. They believe, also, that most persons can be motivated to perform at their best levels when taking a *group* test.

In any testing of groups this must be reasonably well assured, as well as assumed.⁵

Although, as stated, intelligence tests are not designed to measure a person's emotional status and other nonintellectual aspects of personality, there is at present a trend, especially among clinical psychologists, to analyze test performance for evidence of emotional states, personality "dynamics," and for "differential diagnosis" (that is, for symptoms of neuroses, psychoses, or other atypical states). This trend, and the basic idea, while still in a rather early experimental stage, are not without merit and some validity. This aspect of test interpretation, which demands sensitive clinical insights, will be discussed in a later chapter under scatter analysis and under projective analyses of test behavior.

IMPLICATIONS FOR TEST DESIGN AND CONTENT

Definitions of intelligence are of more than theoretical importance. The conception of intelligence which a psychologist holds will affect, to some extent at least, the content of the test he develops. Yet, at the same time, an examination of a representative group of tests reveals the fact that although some are different from others in certain aspects, they all, nevertheless, do have much in common. It would be incorrect to say, for instance, that certain tests exemplify exclusively the definition that intelligence is the capacity to learn. The fact, then, that psychologists emerge with tests having considerable similarity, though they might start with different definitions, must mean that their definitions differ largely in respect to emphases and that, as already pointed out, they are interdependent.

Early experimenters in mental testing attempted to measure general intellectual capacity by means of a single type of test which measured only a single capacity, usually a sensory process, or association, or attention. Thus they identified general intellectual capacity with a

⁵ The "catch" lies, however, in the fact that in any large group of persons being tested it is not unlikely that there will be a few, at least, who are not adequately motivated or who are handicapped by emotional difficulties. Herein is a source of error in group measurement. The discovery of these nonmotivated or blocked individuals will depend upon whether or not each one's rating on the test is scrutinized and evaluated in the light of other and perhaps conflicting evidence. Where there is reason to believe that group-test performance is spuriously low in the case of a given individual, it is desirable to re-examine, using an individual rather than a group test.

single function. Their efforts were fruitless. Later, however, experiment showed that a *variety of test materials* yielded more accurate and more useful results when validated against accepted criteria of intelligent activity. Psychologists, in seeking to encompass a greater variety of items in their tests, and thus to produce more useful and successful instruments, regardless of the exact definition with which each one started out, found that inevitably their testing instruments were broader than their definitions. Current tests thus have more than a little in common in spite of differences in details of their content. Inspection of their content will show that in varying degrees they are measures of some aspects of learning in what is assumed to be a reasonably uniform environment for all persons,⁶ that novel situations and problems are presented, and that ability to carry on abstract thinking is tested through the utilization of symbols and ideas. Inspection will demonstrate also that most of the tests fail to meet the more comprehensive and long-term attributes suggested in Stoddard's definition.

So far as available tests are concerned, it is important to bear in mind that the French psychologist Binet (the "father" of modern mental testing) took the position that it made little difference what specific tasks and items were incorporated into a test provided that in some degree each part was a measure of the individual's general capacity. Whether or not this condition is met will depend, of course, upon the definition of intelligence regarded as most adequate by the designer of a test and upon the criteria of intelligent activity against which test results are checked for validity. Developments have been such that in spite of some differences in definition and in spite of some differences in external appearances, psychologists believe that their tests are reasonably sound because they are related to and have value in predicting the likelihood of intelligent activity in life-situations.

THREE "KINDS" OF INTELLIGENCE

Some psychologists believe that several kinds of intelligence should be distinguished from one another. Noteworthy among them is E. L. Thorndike who has divided intelligent activity into three types: namely, (1) social intelligence, or ability to understand and deal with persons; (2) concrete intelligence, or ability to understand and deal

⁶ This raises the much-debated problem of heredity and environment as factors in the development of intelligence.

with things, as in skilled trades and in working with the appliances of science; (3) abstract intelligence, or ability to understand and deal with verbal and mathematical symbols.

The merit of this classification of types of intelligent activity, for psychological testing and diagnosis, lies in the fact that it indicates several realms in which persons might be functioning and implies that separate and sufficiently specialized tests might be devised to measure how effectively persons are functioning in each.

While it is true that in the case of any given person the scores attained on a test of ability to deal with verbal and numerical abstractions might differ appreciably from those attained by him on a test of social relationships and insights, or on one of "concrete" intelligence, it is true, nevertheless, that when a *representative group* of individuals are tested, the correlations between the types of tests are found to be positive and significant, both statistically and psychologically. For example, correlations between tests of verbal and of concrete abilities vary from about .25 to about .45, the average being about .30-.35. While this is a somewhat low average correlation, it still indicates that some communality of function is being measured. This index, being so far from unity, also means that there are numerous individuals whose relative scores do not correspond closely or whose two relative scores may be discrepant. This fact points up the important psychological principle that the data and status of any single person may be inconsistent with the *general trend*. Study of the individual, and the ways in which and the reasons why he deviates from or exemplifies general trends, is one concern of the clinical psychologist.

Of the three kinds of abilities enumerated above, abstract intelligence is the one that receives greatest weight and is most pronounced in current tests of intelligence—that is, whenever the test is designed for use with persons who are presumed to have reached a level where they reasonably may be expected to have developed facility in dealing with concepts and symbols.

Even tests that present the subject with "things" rather than with ideas and symbols are not devoid of demands upon ability to conceptualize and make abstractions, although testees need not necessarily state these in the form of language and number. For example, when a subject is required to arrange a series of pictures into a sequential and meaningful whole, he must at some stage form a concept of "the whole" if his response is to be correct by some other means than pure

chance. He must do this, also, in assembling parts into an integrated unit (called "object assembly"). The same is true of the child who is asked which is the "prettiest" of two pictures ("aesthetic comparison"), for he must have a concept of "prettiness," however unarticulated it may be. There are in use many other types of test items that deal with things but still require more or less ability in concept formation. Among these are—to name a few—object classification, tracing the shorter of two routes in a maze, identifying objects by use, supplying missing parts in the drawing of a human figure, etc. In short, the fact that some types of test items do not employ language or number does not necessarily signify that they make no demands upon ability to reason at a level of concept formation and abstraction.

It is true that at the earliest developmental levels there are tasks that depend upon visual-motor skill, such as tying a bow knot, grasping a ring, holding a pencil and scribbling, manipulating cubes, and the like. These types of tests, however, are restricted principally to the first eighteen months of life. They are useful as developmental indicators, but they have only slight predictive value for later development of mental abilities, as measured by tests at more advanced levels.⁷

The role of ability to deal with ideas and symbols (words and numbers) as a measure of concept formation and abstraction is of increasing importance in tests of general ability (intelligence) as age level increases. Proportions of verbal and numerical tests, on the one hand, and nonverbal, nonnumerical, on the other, undergo change at different age levels;⁸ some tests include a larger proportion of the latter than do others, even at the adolescent and adult levels.⁹ These differences are not haphazard, nor are they matters of individual whim; they depend upon the purposes of the test and the test author's conception of intelligence and its constituent parts. It will be seen in later chapters that the correlations between various tests of general ability are quite marked—and at times high or very high—thus indicating that to an appreciable degree in these tests the verbal and

⁷ See Chapter 8, "Scales for Infants and Preschool Children," in which this problem is discussed at length.

⁸ Cf. the Revised Stanford-Binet Scale below the age-5 level. Also the Merrill-Palmer Scale, the Minnesota Preschool Test, the Detroit Kindergarten Test, etc.

⁹ Cf. the Dearborn Group tests, the Otis tests, the Kuhlmann-Anderson tests, the Wechsler-Bellevue Intelligence Tests, the Pintner-Paterson scale.

numerical items on the one hand and the nonverbal, nonnumerical on the other are measuring the same or closely related functions.¹⁰ High intercorrelations do not always mean that the same functions are being measured by the tests concerned; such correlations may reflect other common factors which affect the tests being correlated. This, however, is quite improbable as an explanation of test intercorrelations.

✓ ANALYSES OF MENTAL ABILITY

The definitions of intelligence thus far discussed are functional in character; that is, they state how intelligence operates: through learning, adaptation, abstract thinking. But, in addition, psychologists have been concerned to know the fundamental nature and structure of intelligence. They have made analyses in an effort to determine its underlying *factors*. Or, otherwise stated, the purpose of these analyses has been to discover, if possible, the elements, or components, of intelligence, not only for a better theoretical understanding of this complex process, but also to learn what might be the implications for the design and construction of mental tests.

It is not to be inferred, however, that the dynamics of intelligent activity can be adequately understood merely by enumerating and characterizing the components, whatever they might be. Whatever the components, they do not operate independently or in isolation. Understanding the dynamic aspects of mental activity requires some means of characterizing the organization of factors, their interrelationships, and their relation to motivational forces.

Essentially, the experimental method followed is this: a rather large number of separate tests, more or less diverse in character, are given to an adequate sampling of the population. The results of each type of test are correlated with those of all the others. The coefficients of correlation are then subjected to various techniques of statistical analysis in an effort to discover the extent of common ground between them (technically known as *communality*), and their degree of independence. These statistical methods are known as *factor analysis*.¹¹ The particular theory or structure of intelligence deduced from the statistical operations will depend upon the expert's own interpretation of

¹⁰ The statements in this paragraph do not mean that the nonverbal materials in the tests are measures of mechanical ability. Generally, they are believed by the authors of the tests to measure the same psychological processes as do the verbal materials, but by means of different content.

¹¹ To be discussed later in this chapter.

the analysis;¹² and the experts differ in their interpretations. These differences, however, need not invalidate the use of well-standardized psychological tests; for, as will be seen, theoretical differences thus far have not had far-reaching consequences as regards the kinds of intelligence tests constructed.

The Multi-factor Theory. Thorndike's multi-factor theory of intelligence is at one extreme of the interpretations regarding the nature of mental organization. As the name of the theory indicates, intelligence is said to be constituted of a multitude of separate factors, or elements, each one being a minute element of ability. Any mental act, according to this theory, involves a number of these minute elements operating together. Any other mental act also involves a number of the elements in combination. Hence, if performances on these two mental tasks are positively correlated, the degree of correlation is due to the number of *common elements* involved in the two acts. If two types of mental activities, A and B, are more highly correlated than are A and C, the reason, according to the multi-factor theory, would be that the first pair has more elements in common than does the second pair. According to this theory, then, there is really no such factor as "general intelligence"; there are only many highly specific acts, the number of such depending upon how refined a classification we might wish to make and are capable of making.

Thorndike's is really an "atomistic" theory of mental ability. He adds, however, that certain mental activities have so many of their elements in common that it is useful to classify these tasks into separate groups to which special names are given; for example, verbal meaning, arithmetical reasoning, comprehension, visual perception of relationships, and others. Consequently, in constructing a mental test, it appears even to Thorndike himself that his "atomistic" theory and the multitude of minute elements of ability are of less practical significance than the conception that many of them operate together in any situation demanding intelligence. This is illustrated by Thorndike's test designed to measure ability to deal with abstractions. His test is composed of four parts: sentence completion (C), arithmetical reasoning (A), vocabulary (V), and following directions (D). This instrument is known as the *CAVD test*. It is not claimed by Thorndike that these four sets of items encompass the entire range of ab-

¹² This, of course, is true of all sciences.

stract intelligence. They represent and sample only certain parts; but because of the very significant correlations between all types of measures within the tested range, it is held, the other aspects of abstract intelligence can be estimated with satisfactory accuracy from those portions that are actually measured by this test.

The Two-factor Theory. Opposed to Thorndike's theory of the nature of intelligence is Spearman's two-factor theory, which stands at the other extreme of interpretations. According to Spearman, all intellectual activity is dependent primarily upon and is an expression of a *general factor* common to all mental activity. This factor, designated by the symbol *g*, is possessed by all individuals, but in varying degrees, of course, since people differ in mental ability; and it (*g*) operates in all mental activity, though in varying amounts, since mental tasks differ in respect to their demands upon general intelligence. Spearman characterized this general factor as mental energy, because in the realm of intelligent activity, he maintained, it has a role similar to that of physical energy in the physical world. Like all other scientific concepts, the general factor can be observed and known only through its specific manifestations—in this instance, through psychological tests. After analyzing tests with varying amounts of the general factor, from high to low, Spearman concluded that the principal distinguishing characteristic of tests highly "loaded" with *g* is that they require insight into relationships—what he called "the eduction of relations and correlates." For example, in solving an arithmetical problem, the subject has to grasp the relationships between the data presented, organize them with reference to the propositions given in the problem, and deduce a correct answer. The *g*-content in this task is high. By contrast, if the subject merely has to repeat a table of multiplications or add a few numbers—both of which can be learned by rote—no insights are necessary and no relationships need be grasped. In this task, the amount of *g* involved is very small.¹³

Spearman postulated the *g* factor, in the first place, to explain correlations that he found to exist among diverse sorts of perceiving,

¹³ When using an individual test, like the Stanford-Binet or the Bellevue, it has often been observed by examiners that a subject who is unable to solve an arithmetical *problem*—unable even to make a start toward a solution—may still be able to perform the separate arithmetical *processes* involved. In Spearman's terms, such an individual is unable to educe the necessary relations and correlates, for lack of the necessary amount of *g*.

knowing, reasoning, and thinking, as illustrated in Table 9. That is to say, he concluded that all mental activity is to some extent dependent upon and an expression of this general factor; and the magnitude of the correlation coefficient found between any two forms of mental activity reveals the extent to which this *g* factor is operative in each and common to both. Thus, the amount of the general factor operating in each activity will determine the size of the correlation between the two mental activities being measured. The types of materials used in current tests of intelligence—word meaning, arithmetical reasoning, sentence completion, reasoning by analogy, paragraph interpretation,

TABLE 9
Intercorrelations of Subtests¹⁴

	1	2	3	4	5	6	7
(1) Analogies	.50	.49	.55	.49	.45	.45	
(2) Completion	.50	...	.54	.47	.50	.38	.34
(3) Understanding paragraphs	.49	.54	...	.49	.39	.44	.35
(4) Opposites	.55	.47	.49	...	.41	.32	.35
(5) Instructions	.49	.50	.39	.41	...	.32	.40
(6) Resemblances	.45	.38	.44	.32	.32	...	.35
(7) Inferences	.45	.34	.35	.35	.40	.35	...

perception of relationships in geometric forms, picture completion, and others—all show significant degrees of positive correlation with one another. Spearman and his supporters at first ascribed this fact to the presence of *g*, in greater or lesser amount, in all of them. Later researches led them to conclude that certain “group factors” are also present in some mental activities. These are the factors that occur in more than one type of test item, but in less than all of any given set of tests. The general factor, however, still remains the primary and pervasive one.

Since the intercorrelations are by no means perfect, Spearman postulated the existence of specific factors, called *s* factors, each of which is specific to a particular type of activity. Thus, the two-factor theory states that all mental activities have in common some of the general factor; each mental activity might also be a member of a “group”; and each has also its own specific factor. Of the kinds of factors, the general one is regarded as the essential measure of intelli-

¹⁴ From C. Spearman, *The Abilities of Man*, New York: Macmillan, 1927, p. 149. (By permission.)

gence; accordingly a sound test of intelligence is one that will sample adequately the *g* factor in a variety of activities, and the best test materials are those which call for the largest amount of the general factor. And the largest amounts of the general factor are believed to be demanded by those types of test materials that have the higher inter-correlations with one another.

As a matter of fact, since the beginning of modern mental testing, psychologists have proceeded, at least implicitly, on the assumption that all forms of mental activity have something in common—that they are similar in certain basic respects. Otherwise, psychologists could not have justified their practice of testing together in a single instrument such diverse mental activities as defining words, solving arithmetical problems, finding similarities and differences, repeating digits forward and backward, completing sentences in a meaningful manner, perceiving geometric forms, etc. All of these, and the others used, must have been regarded as being measures, to a greater or lesser degree, of general intelligence. From the total performance on these tests, it was believed that an individual's level of general intelligence would emerge. *Therefore, psychologists believed they were justified in adding up the test items correctly passed in the several types of activity and deriving a single total score to represent an individual's general intelligence level.*¹⁵ This is the actual practice being followed in nearly all tests, including individual as well as group scales of mental ability.¹⁶

The practical implication of the Spearman two-factor theory is clear, so far as test construction is concerned. A test conforming to this theory would be one whose materials and several parts are saturated with the general factor so that measurement thereby would cause

¹⁵ While this practice is not being discontinued, and should not be, increased emphasis is now being placed on the desirability of representing each individual by means of a test profile, where possible, as well as by a general index. There are some psychologists, however, who would abandon the use of all indexes of general level and would substitute a profile representing the individual's relative rank in each of the specific types of test materials being used; e.g., numerical ability, word meaning, spatial perception, and the like.

¹⁶ In addition to *g* and *s*, Spearman and others have found by further analysis of experimental results that there are some nonintellectual factors—such as volition, interest, persistence—which influence a person's effectiveness. Spearman and adherents of his theory have also discerned a few *groups* of factors that are intermediate between *g* and the highly specific *s*. They suggest that musical aptitude and mechanical aptitude are of this type.

the testee's level and quality of g to emerge, while the effects of specific factors (s) would be canceled out. Thus, the net result of the test would be a measure of g . To achieve this would require a skillful selection and development of test problems and parts that are significantly intercorrelated, which at the same time satisfy the practical criteria of intelligent activity. Such a test, presumably, would yield an index which reflects the caliber of a particular mentality working as a whole.

The Group-factor Theory. Intermediate between the theories of Thorndike and Spearman are the group-factor theories, from among which we select for presentation that of Thurstone because it has been most highly developed, has received most consideration, and has resulted in the construction of a set of measures called *tests of primary mental abilities*.

According to the group-factor theory, intelligent activity is not an expression of innumerable highly specific factors, as Thorndike claimed. Nor is it the expression primarily of a general factor which pervades all mental activity and is the essence of intelligence, as Spearman held. Instead, the analyses and interpretations of Thurstone and others led them to the conclusion that *certain* mental operations have in common a "primary" factor which gives them psychological and functional unity and which differentiates them from other mental operations. These mental operations, then, constitute a "group." A second group of mental operations has its own unifying "primary" factor; a third group has a third; and so on. In other words, there are a number of groups of mental abilities (the number being as yet undetermined) each of which has its own "primary" factor, giving the group a functional unity and cohesiveness. Each of these "primary" factors is said to be relatively independent of the others.

After administering a large variety of types of test materials to college students and to high-school and eighth-grade pupils, and after making correlational analyses of the results, Thurstone and his collaborators concluded that six "primary" factors emerged clearly enough for identification and use in test design and construction. They are, briefly, the following.¹⁷

¹⁷ From L. L. Thurstone and T. G. Thurstone, *The Chicago Tests of Primary Mental Abilities*, Manual of Instructions, Chicago: Science Research Associates, 1943, p. 7. See also L. L. Thurstone, *Primary Mental Abilities*, Psychometric Monograph No. 1, Chicago: University of Chicago Press, 1938; L. L. Thurstone

The Number factor (N): "ability to do numerical calculations rapidly and accurately."

The Verbal factor (V): "found in tests involving verbal comprehension."

The Space factor (S): "involved in any tasks in which the subject manipulates an object imaginably in space."

The Word Fluency factor (W): "involved whenever the subject is asked to think of isolated words at a rapid rate."

The Reasoning factor (R): "found in tasks that require the subject to discover a rule or principle involved in Series or groups of letters." Although it is believed both induction and deduction are involved, it seems that induction is the more significant here.

The Rote Memory factor (M): involving "the ability to memorize quickly."

In spite of the fact that "primary" mental abilities (or factors) were originally said to be functionally independent of each other, it was actually found that they are positively and significantly intercor-

TABLE 10
*Intercorrelations of Subtests*¹⁸

	N	W	V	S	M	R
N	...					
W	.41	...				
V	.40	.54	...			
S	.28	.17	.16	...		
M	.31	.36	.35	.13	...	
R	.53	.49	.59	.29	.39	...

(N, number facility; W, word fluency; V, verbal meaning; S, spatial perception; M, rote memory; R, reasoning)

related, as shown in Table 10. This must mean that the "primary" and presumably independent factors are not the only factors at work in the mental activities required by the tests. There must be some other

and T. G. Thurstone, *Factorial Studies of Intelligence*, Psychometric Monograph No. 2, Chicago: University of Chicago Press, 1941. Some modifications of factors have been introduced in the most recent issues of these tests for younger subjects. The six named above are still regarded as established factors.

The "primary" mental abilities do not include the entire range of human abilities. They do not include mechanical, musical, or artistic aptitudes. The "primary" abilities involve largely, though not entirely, those which are needed in "abstract intelligence" and in the academic types of learning.

¹⁸ From L. L. Thurstone and T. G. Thurstone, *The Chicago Tests of Primary Mental Abilities*, Manual of Instructions, Chicago: Science Research Associates, 1943. (By permission.)

factor, or factors, to account for the common ground (as shown by the positive correlations) that exists between the various psychological tests intended to measure these "primary" factors. In other words, it seems that the test authors have not been able to devise test materials which will sample the "primary" mental abilities in "pure" form. The Thurstones, therefore, concluded that in addition to the "primary" abilities there is a "second-order general factor." They also stated, in their earlier test manual, that "If further studies of the primary mental abilities should reveal this general factor, it may sustain Spearman's intellectual factor."¹⁹

Subsequent studies of the "primary mental abilities" do tend to reveal a general factor. The more recent intercorrelations found among the several tests that make up the PMA²⁰ batteries are quite marked, especially at the lower age levels (when abilities are less differentiated through education and interest than they will be in later years).

For the tests at the 5 to 7 year level, the intercorrelations range from .46 to .67 (average equals .55). For tests at the 7 to 11 year level, the range of coefficients is from .41 to .70 (average equals .50); while for ages 11 to 17, the range of coefficients is from .13 to .59 (average equals .30+). It appears, thus, that the group-factor adherents have found it necessary to posit the operations of a general factor; but at present they regard it as being of a "second order."

In evaluating the group-factor hypothesis we need not question the soundness of the statistical methods used or the comprehensiveness of the experimentation.²¹ Several observations are necessary, however, to enable the reader to make a fuller assessment of the hypotheses. In the first place, intelligence is not an entity which operates in a vacuum; it is not something "given," even in the sense that some physical traits are "given," such as the color of eyes and hair, the number of digits, etc. Intelligence is, rather, a name for certain kinds of activity; we can know of it only through its manifestations in behavior. Intelligent behavior develops and is manifested in one kind of environment or another; hence, the particular form of expression that intelligent activity takes will depend upon the sort of functions which are developed and fostered in a given cultural environment. In our own and similar cul-

¹⁹ *Ibid.*, p. 7.

²⁰ Primary Mental Abilities.

²¹ Some psychologists have criticized adversely Thurstone's methods and his interpretations of his findings.

tures, verbal and numerical abilities are, of course, essential; they are fostered from earliest childhood, and they receive greatest attention and emphasis in our schools. There is thus a relationship between this cultural emphasis and the fact that three of Thurstone's six "primary" factors are concerned with words and numbers. It is probable also that the "Space" factor emerges from statistical analyses because of our experiences with things in three-dimensional space. The two remaining factors, "Reasoning" and "Rote Memory," are characteristic, in greater or lesser degree, of all persons regardless of the particular culture; we should therefore expect to find them as factors of intelligent activity in any analysis. Furthermore, the "Reasoning" factor is very likely much the same as Spearman's *g*, although the latter has been presented as having several aspects.²² In effect, then, the point is that some of the *particular* factors through which intelligence is expressed are developed by experience and education; and these particular factors—such as the six "primaries"—may well be conceived of as particular manifestations of a *general* ability rather than as "primary" abilities.

The proponents of the group-factor hypothesis do not claim that the exact number of "primary" mental abilities is known. Hence, we must caution the reader against assuming that there is a finality about the present number. For example, a factor of "Speed" will not appear in a statistical analysis of test results unless speed of performance is, first, a variant in the population being measured and, second, a requirement within the tests themselves. The same can be said of "persistence" or "mental fatigue." Similarly, "originality" would appear as another factor, *if* it could be measured.

These observations do not invalidate the tests that have been or might be designed and constructed on the basis of the group-factor hypothesis. As a matter of fact, the contents of such tests thus far published, though differently organized, are not radically different in their essentials from those that have been designed and constructed on the basis of either of the two other theories of the nature of intelligence.

For our present purposes, two consequences of group-factor analyses are indicated. First, the conceptual framework has resulted in more clearly specified and defined test categories and types of test items

²² Apprehension of one's own experience, the education of relations, and the education of correlates.

than was the case previously. Second, several batteries of tests have been constructed on the basis of group-factor theory, particularly the Thurstones' tests of *Primary Mental Abilities*.

The early versions of the PMA tests did not yield an over-all index of performance, such as mental age, intelligence quotient, or an over-all percentile rank. Instead, they gave for each subject separate percentile ranks to represent his performance level in each of the "primary" factors. These ranks were then used to make, for each person, a "mental profile" for purposes of educational and vocational guidance. While they granted that the single index (mental age and IQ), based upon a variety of mental activities, is useful for many practical purposes, group-factorists originally maintained that their method of finding separate ratings for each of the "primary" factors enables the examiner more readily and adequately to recognize a testee's marked mental abilities and disabilities, the degree of uniformity or lack of uniformity.²³

There is merit in this contention; yet, at the same time, there is no good reason why the group-factor type of test should not also yield an over-all rating (such as MA or IQ) as well as indexes of relative rank for each of the specified factors. While mental age and intelligence quotient should never be interpreted and used mechanically and uncritically, or in disregard of the specific performances that have contributed to them, they do nevertheless have considerable significance and valuable connotations for the qualified examiner and interpreter.

The group factorists have apparently recognized this point in their more recent interpretations of test findings. Equally important to them—if not more so—is the fact that their correlations and factorial analyses have persistently yielded results that could not be explained in terms of group factors alone, and that the *g* factor was indicated. As a result, the most recent editions of the PMA tests provide IQ equiva-

²³ This can be done by a competent examiner when another group test or the Binet or the Bellevue is used. In the case of the last two the "scatter" of test performance is studied. In the case of group tests, the individual's performance on each of the several parts can be compared with his performances on the other parts. The disadvantage here lies in the fact that the usual group test does not provide separate scores and norms for each of the parts.

A weakness of the group-factor type of test is this: the breakdown into separate factors ignores the fact that intelligence expresses itself in behavior as a combination, a unity of functions, not as a series of independent factors.

lents for the scores on the scale for ages 11 to 17, while for the younger age levels both MA units and quotients are provided.

As is so often the case in scientific problems—especially in the relatively new ones—divergent theories in time tend to come into closer agreement. The Spearman Two-Factor Theory now recognizes that some group factors should be posited to explain test findings; but emphasis is upon the *g* factor. Perhaps the Spearman theory may now be renamed “The General Factor-Group Factor Theory,” while the other might be renamed “The Group Factor-General Factor Theory.” The narrowing of differences between the two theories represents significant scientific progress.

FACTOR ANALYSIS

The two-factor and the group-factor theories are the two most prominent examples of doctrines emerging from the methods of factor analysis. Although this subject is highly technical, it is desirable to explain it somewhat more fully at this stage.

The technique is essentially a search for the psychological functions which are at the basis of and determine test performance. All techniques of factor analysis are statistical and based upon the correlation coefficient. After the statistical calculations have been made, it is necessary for the investigator to bring to bear his psychological insights to interpret and name his statistical findings. Tests contain a variety of items. What psychological functions do the various types of items have in common? Are there functions in common between various tests of verbal performance? Between verbal and numerical? Between spatial perception and numerical ability? Between reasoning with verbal and with nonverbal materials? These are among the questions the factor analyst seeks to answer. After he has found his answer, at least tentatively, he proceeds to construct a scale in which items are included and so grouped as to measure only, or almost solely, the factors he has segregated from his preliminary testing and statistical analysis.

The factor analyst does not begin with a definite set of preconceived mental functions. He tries to discover which psychological functions, or components, are necessary to explain his data. Yet, it should be noted, he must at the very outset have some conception of the kinds of test items to include in preliminary experimentation. Thus what he ultimately distills out as factors is basically dependent upon his origi-

nal conceptions regarding his preliminary items. The factor analyst, in seeking the components of "intelligence," for example, does not start with tests of color perception, tone discrimination, or finger dexterity.

Two-factor Theory. We have already stated Spearman's *two-factor theory*. It will be helpful now to describe in more detail the reasoning whereby the theory was arrived at. Spearman, in his early experimentation, was impressed by the fact that all the intercorrelations were positive in a table of coefficients for various types of items. He was also impressed by what appeared to be a hierarchy of coefficients in the rows and columns of the table; not perfect gradations, but strong evidence of proportional gradations.²⁴ He therefore offered a hypothetically perfect table of correlation coefficients to illustrate his point (Table 11).

TABLE 11
Spearman's Hypothetical Table of Correlations ²⁴

	1	2	3	4	5
(1) Opposites		.80	.60	.30	.30
(2) Completion	.80		.48	.24	.24
(3) Memory	.60	.48		.18	.18
(4) Discrimination	.30	.24	.18		.09
(5) Cancellation	.30	.24	.18	.09	

Not only are the coefficients positive and in a decreasing order along rows and columns, but theoretically any two columns (or rows, since the table is symmetrical about the diagonal which contains the self-correlations) are in direct proportion. The criterion of proportionality requires that the following correlational relationships should hold:

$$\frac{r_{13}}{r_{23}} = \frac{r_{14}}{r_{24}} = \frac{r_{15}}{r_{25}}.$$

Taking only the first two ratios,

$$\frac{r_{13}}{r_{23}} = \frac{r_{14}}{r_{24}},$$

and multiplying by the demoninators, we have

$$r_{13} r_{24} = r_{23} r_{14}.$$

²⁴ From C. Spearman, *The Abilities of Man*, New York: Macmillan, 1927, p. 74. (By permission.)

Transposing, we get

$$r_{13} r_{24} - r_{23} r_{14} = 0.$$

From this *tetrad equation* (so called because the test correlations are dealt with in sets of four) may be obtained what is known as the *tetrad difference*.

The tetrad equation may be written for the combination of any four tests. By rearranging the four coefficients, three tetrad differences may be obtained for every combination of four tests. Thus, using t as the notation for tetrad difference:

$$t_{1234} = r_{12} r_{34} - r_{13} r_{24}$$

$$t_{1243} = r_{12} r_{34} - r_{14} r_{23}$$

$$t_{1342} = r_{13} r_{24} - r_{14} r_{23}$$

Theoretically, the tetrad difference criterion is satisfied if t is zero. When it is zero, Spearman and others have offered mathematical evidence to demonstrate that a single common factor can account for the relationships among the four tests, or variables. But in actual fact, the difference is rarely if ever zero. However, if the differences are close to zero, we may also conclude the criterion is satisfied, since correlations between tests will be affected by errors of measurement due to chance and accidental factors.²⁵ (Cf. discussion of reliability in Chapter 1.) Furthermore, the correlation coefficients would also be affected by the operations of the *specific* factor in each test. The specific factor was postulated to explain, in part at least, the tetrad differences that were greater than zero.

If more than four tests are being examined to disclose the functions involved, we may substitute, say, tests numbers 5 and 6 in the tetrad equations in place of numbers 3 and 4. Thus we would be analyzing tests 1, 2, 5, 6. Then if the tetrad difference criterion is satisfied, it would be concluded that the functions common to 1 and 2 are identical with those common to 5 and 6. Assuming that the tetrad difference criterion was satisfied also for tests 1, 2, 3, 4, then the same functions (or factor) are said to be common to the six variables. The same reasoning may be applied to any number of tests.

The principal conclusion drawn by Spearman and some others, after their analyses, was that the degree of correlation between any

²⁵ Formulas are provided for calculating probable errors of tetrad differences; comparison of a tetrad difference with its probable error (PE) enables one to decide whether it differs significantly from zero.

Factor Analysis

two tests is dependent upon the extent to which g is involved in each. Subsequent investigations showed, however, that *some* test intercorrelations may include their own common factors beyond the single common g . It was necessary, therefore, to postulate the operations of *group factors*, each group being effective in two or more tests, but not in all of them. Spearman and others recognized such group factors as numerical, verbal, speed, mechanical, imagination, and attention. In addition, Spearman had postulated three nonintellective factors which influence one's mental effectiveness: perseveration (p); oscillation (o), being one's variability in performance in continuous mental activity; and will (w), being one's persistence in effort.

The tetrad difference criterion does not in itself show the relative weight or importance of the common factor in each kind of test. Following the work of Spearman, therefore, methods have been developed for finding the weights (commonly called "loadings") of each factor—general or group—in each of the intercorrelated tests. These methods are known as "factor pattern analysis."²⁶

The two-factor theory can account for the universal positive correlation coefficients among the various kinds of test items included in scales to measure mental ability, since every form of test requires the operation of g to some degree. Pooling a variety of kinds of tests in a scale is sound practice, according to this theory, because we thereby approximate a measure of pure g . Since the s factors are uncorrelated within any individual—that is, they may be possessed in varying and random degrees by him—they will be a negligible factor in the total performance in a pooled test of general ability, because the varied s factors will tend to cancel out one another.

Sampling Theory. The two-factor theory has been criticized by some statistical psychologists, notably G. H. Thomson and L. L. Thurstone. Thomson offers a *sampling theory*²⁷ to explain the same tables of intercorrelations. Briefly, his view is that the coefficients of correlation are the results of common samplings and combinations of independent factors. The number of common independent factors utilized by two

²⁶ For example, T. L. Kelley, *Essential Traits of Mental Life*, Cambridge: Harvard University Press, 1935; L. L. Thurstone, *Vectors of the Mind: Multiple-Factor Analysis for the Isolation of Primary Traits*, Chicago: University of Chicago Press, 1935.

²⁷ G. H. Thomson, *The Factorial Analysis of Human Ability*, Boston: Houghton Mifflin, 1939, Chapter 3.

tests will determine the coefficient of correlation between these two. This theory is, of course, the same as Thorndike's, except that Thomson concedes the practical usefulness of a concept like *g*. Thomson also adds that if several tests call upon many elementary factors in common, they will not only have a very marked or high coefficient of correlation, but they will give the appearance of having *one* common comprehensive factor. Also, Thomson's theory maintains that if several tests draw upon a relatively smaller number of the elementary factors in common, we have then group factors—that is, a limited number of factors that enter into performance on types of tests which are distinguished by the fact that they have certain mental processes in common but do not share a very large number of elementary factors or a universal *g*.

While both theories require that a scale to measure general mental ability should pool a variety of types of tests differing in content and mental processes employed, the two-factor theory would seek subtests (parts of the scale) that have high intercorrelations, whereas the sampling theory would seek subtests having low intercorrelations among themselves but high correlations with the criteria of validity.

Group-factor Theory. As already stated, Thurstone and others believe that a *group-factor theory* fits the facts best and is most useful in testing practice. Their view differs from Thomson's in that they reject the theory of a very large number of independent factors. As previously explained, a group-factor is conceived of as an operational concept to account for correlations of performance within only a limited group of tests.²⁸ Several different groups of factors are necessary to account for all mental activity, plus the more recently added second-order *g* factor, which Thurstone states may be more "central" in character and more "universal" in influence.

Thurstone has contributed very significantly to the methodology of group-factor analysis, particularly his geometric methods; and from his analyses, as we shall see, has emerged a scale to test mental ability. This volume is not the place to present his techniques; we shall merely state his purposes.²⁹

²⁸ Most recently, some group-factor theorists have characterized "primary factors" as facilities of the mind and as media of expression.

²⁹ A number of others have made significant contributions to factor theory, especially K. J. Holzinger, H. Hotelling, R. C. Tryon, H. E. Garrett, C. L. Burt, J. C. Flanagan, J. P. Guilford, and P. Vernon.

Three objectives, according to Thurstone, are to be achieved by factor analysis: (1) determination of the smallest number of primary mental abilities to be postulated as an explanation of tables of inter-correlations; (2) determination of the amount of each primary ability that is involved in each test; and (3) determination of regression

TABLE 12

The Two-Factor Pattern

Test	General Factor	Specific Factors
1	x	S ₁
2	x	S ₂
3	x	S ₃
4	x	S ₄
5	x	S ₅
6	x	S ₆

The Group-Factor Pattern

Test	Group Factors			
	A	B	C	D
1	x		x	x
2	x	x	x	
3		x		x
4	x	x	x	
5	x	x		x
6		x	x	

Factor Theories Combined

Test	General Factor	Group Factors			Specific Factors
		A	B	C	
1	x	x			S ₁
2	x	x			S ₂
3	x		x		S ₃
4	x		x		S ₄
5	x			x	S ₅
6	x			x	S ₆

equations whereby the amount of a primary mental ability in an individual can be estimated from tests that draw upon that ability. As an illustration, we may consider several tests in which only two group factors are involved. An individual might make a high score in these tests either by having a moderately high level of ability in each of Factor I and Factor II, or by having very much of one and little of the other. Also, if Factor I carries much heavier weight in the tests

than does II, then a high level of ability on I is more important for high performance on these tests than is a high level of II. Thus, the Thurstone method would find the relatively few primary or basic mental abilities, devise a scale to measure all of them, and so organize and score the subtests as to reveal each individual's relative strength in each factor.

Summary. Methods of factor analysis differ somewhat in their assumptions, and analysts differ somewhat in their interpretations or results, but the general conclusions derived by the several methods of analysis and interpretation do not differ radically. All factorial theories now postulate the presence of group factors, although the groups are not always identical and differ in relative emphasis placed upon them by different theories. Most theories also find a general factor necessary to explain intercorrelations, although here again emphasis upon the general factor varies. All agree that an individual's mental activities are attributable to the various ways in which the general and group factors combine in the performance of varied mental tasks. While several methods of factorial analysis are possible, basically the choice between them and interpretations derived through them must rest upon psychological theory and concepts rather than upon statistical methods. Factors should not be regarded as fixed, predetermined mental entities. The factors that are found are influenced by the ages of the persons tested, by interests, by experience and training, and by the test items originally employed in the preliminary investigations. Factorial analysis is a statistical method which provides the means of improving test construction and of classifying test performance.

ILLUSTRATIONS OF FACTORS

The following illustrations will assist the student to grasp more fully the problem of factors in psychological testing. As already stated, factor analysis techniques depend basically upon test intercorrelations. Tables 13A and 13B show the intercorrelations between two sets of four subtests of the Wechsler scale for children.

Since all the coefficients in Table 13A are positive and quite marked, we conclude that all four tests have much in common, but that Vocabulary, Information, and Similarities have somewhat more in common with one another than they do with Comprehension. But

since all the coefficients are far from perfect (+1.00), we are warranted in using all four to sample the testee's mental abilities, rather than only one or two to the exclusion of the others. The fact that these four tests are not perfectly correlated—nor nearly so—might be due to one of these possibilities: (1) that each test samples the *g* factor in different amounts, plus its own specific factor; or (2) that the tests have *g* in common but each test samples also one or more group factors, though not necessarily the same ones; or (3) that each has many highly specific factors in common with every other one, as

TABLE 13A
*Intercorrelations of Four Subtests of the Wechsler Intelligence Scale for Children*³⁰

	Vocab.	Info.	Sims.	Comp.
Vocabulary	—	.74	.66	.60
Information		—	.67	.61
Similarities			—	.61
Comprehension				—

TABLE 13B
Obj. Assemb. Comp. Arith. Digit Sp.

	Obj. Assemb.	Comp.	Arith.	Digit Sp.
Object Assembly	—	.13	.20	.13
Comprehension		—	.46	.28
Arithmetic			—	.40
Digit Span				—

well as unique factors. A technical factorial analysis would go beyond this inspection analysis in an effort to determine which of these three hypotheses is the most plausible one, and to determine to what extent performance on each of the four tests calls upon whatever factors—*g* or others—might be inferred from the statistical analysis.

Table 13B, by contrast, shows four subtests that have low intercorrelations. Of the six coefficients, only two (.40 and .46) are large enough to suggest that the subtests involved have much in common as regards psychological functioning. The coefficient of .46 between Comprehension and Arithmetic is attributable to the demands that both of these tests make upon reasoning ability, or, more specifically, ability to analyze a set of given material and then reorganize the elements toward the solution of the specified problem. The coefficient

³⁰ From the *Manual*. Psychological Corporation.

of .40 between Arithmetic and Digit Span is attributable, it appears from the characteristics of the two tests, to facility with numbers and ability in immediate recall (as contrasted with delayed recall). The remaining four coefficients are so low as to suggest that the tests concerned have little dependence upon common functions (whether *g* or other factors), and that each makes demands upon some factor or factors not called upon by the others. Here again, a factorial analysis would attempt to identify more precisely the factors involved; but in so doing the analyst would have to apply his knowledge of psychological functioning to the items that are shown by the analysis to cluster together.

Figure 3.1 shows in graphic form how two and three tests *might be* interrelated. As the number of types of tests increases, the possible factor interrelationships may become more numerous and complex, though it is extremely improbable that *no* overlapping whatever of factors would be found in measuring human abilities by means of two or more different types of tests. In view of the more recent partial reconciliation of the group-factor and the general-factor theories, the illustrated overlappings are most probably attributable to *g*.

The possible factor interrelationships of the parts of Figure 3.1 are:

A. Each test is factorially independent of the other. The factor or factors in each are unique to it, either as "group" factors or as "specific" factors.

B. The overlapping shaded area indicates a factor or factors common to both tests. This may be *g* or a group factor. The unshaded area indicates factors unique to each, either specific or group, or both. When a number of diverse tests show some overlapping among all of them, the soundest inference is that a *g* factor accounts for the common ground.

C. In this instance the tests may be measuring only the general factor, or *g* plus the same group factor, or just the same group factor. There are no unique factors. It is extremely improbable, in this situation, that the general factor is not involved. If numerous pairings of different tests showed this relationship, the soundest inference would be that a general factor is being measured.

D. Each of the three tests is factorially independent of the other two. The uniqueness of each may be due to group or specific factors, or to both.

E. The overlapping of 1 and 2 here may be attributable to *g* or to a group factor. Test 3 is independent of the others. The nonoverlap-

ping segments of 1 and 2 may represent separate group factors or specific factors in each test.

F. In this figure, overlapping between 1 and 2, and between 2 and 3, is attributed to one or more group factors, but different ones in each case, since there is no common ground between all three tests.

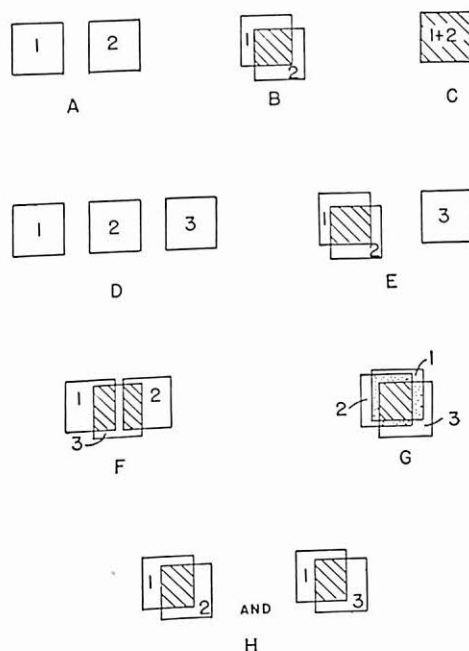


FIG. 3.1. Possible Intercorrelations Between Two and Three Tests.

The unshaded areas would represent special factors or group factors, or both, which are not shared with either of the other two tests.

G. Here there is some common ground in all three tests (shaded area), which is interpreted as showing the presence of the *g* factor. The dotted areas show group factors shared by only two of the tests. The unshaded areas represent either specific factors or unique group factors, or both.

H. This figure represents three tests that have only the general factor in common. Both Tests 2 and 3 have the same amount and the identical area in common with Test 1; hence, they have the same amount and identical area in common with each other.

These graphic illustrations of correlations and factor "loadings" derived from statistical analysis serve three purposes: (1) They demonstrate the complexity of the problem of determining interrelation-

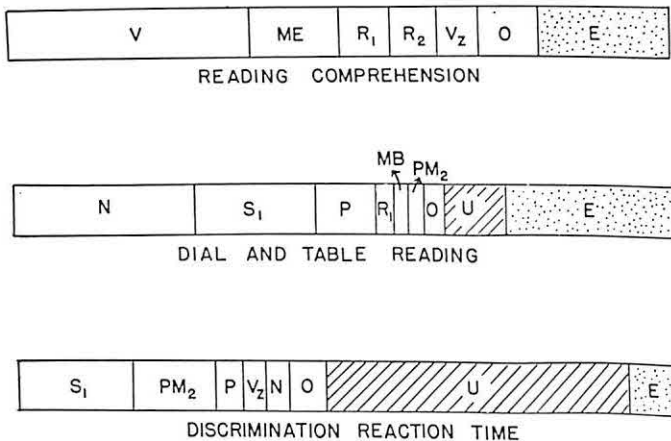


FIG. 3.2. Diagrams of the Component Variances of Three Army Air Force Classification Tests. (From Guilford, op. cit. p. 86.) The letters stand for:

- V verbal-comprehension factor
- ME mechanical-experience factor
- R_1 reasoning I (general-reasoning) factor
- R_2 reasoning II (common to analogies tests) factor
- V_2 visualization factor
- O other common factors, each with variance too small to mention separately
- U unknown common-factor or specific-factor variances
- E error variances
- N numerical factor
- S_1 space I (spatial-relations) factor
- P perceptual-speed factor
- MB mathematical-background factor
- M_2 memory II (visual-memory) factor
- PM_2 psychomotor II (precision) factor

ships of psychological factors. (2) They demonstrate that the same statistical findings are often open to more than one psychological interpretation; and, using the statistical findings as aids, one's interpretation will depend basically upon his psychological analyses of intellectual functioning. (3) The illustrations and their several possible

interpretations help to make clear the reasons why the most valid and useful tests within a given category (e.g., intelligence) have much in common as regards psychological functioning and as regards test items themselves.

Finally, Figures 3.2 and 3.3 illustrate elaborate factorial analyses of tests which have been statistically fractionated.³¹ These indicate the probable quantitative portions of each of the several factors in each of the tests. Such analyses do, undoubtedly, provide insights into the

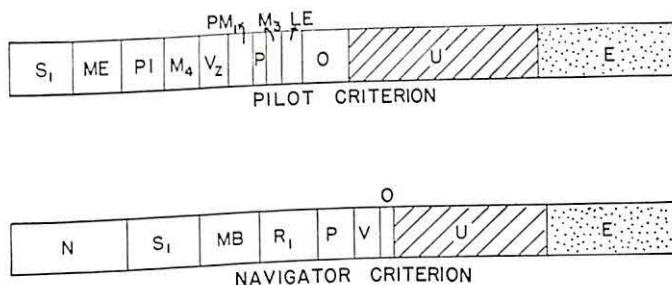


FIG. 3.3. Diagrams of the Component Variances of Pilot and Navigator Training Criteria. (From Guilford, op. cit., p. 86.) Letter symbols as defined with Figure 3.2, except for some additional ones:

- PI pilot-interest factor
- M₄ memory IV (content-memory) factor
- M₃ memory III (picture symbol association) factor
- PM₁ psychomotor I (coordination) factor
- LE length-estimation factor

psychological operations that combine in performance on the tests. Test construction is thereby facilitated. It should not be assumed, however, that each of these factors exists or operates independently. We may look at the factors in the Reading Comprehension test as an example. We note that "verbal comprehension" is the largest single factor; then we have, in order, "mechanical experience," "reasoning I" and "reasoning II." It is very doubtful that these four factors can or should be separated functionally. Mechanical experience will have been significant in the comprehension of the verbal materials of this test and will have contributed to the verbal competence of the testee

³¹ J. P. Guilford, "Factorial Analysis in a Test-Development Program," *Psychological Review*, Vol. 55, 1948, pp. 79-94.

in the particular area covered by the test. Reasoning, of whatever kind, is basically a matter of problem-solving ability, whether with the use of concrete objects or with abstractions (words and numbers). In this test of Reading Comprehension, the examinee's ability to reason with the problems presented will be dependent in part upon his mechanical experience (contributing to his comprehension) and to his knowledge of the terms used in the problems. Conversely, the extent and quality of the vocabulary he acquires and the degree to which he benefits from his mechanical experience will depend in part upon his present and potential reasoning capacity.

IMPLICATIONS

The hypotheses as to the nature of mental abilities have been arrived at by means of several methods of statistical analysis and through partially different interpretations placed upon much the same data by different investigators. Regardless of which of the hypotheses an author of a test follows, the instrument he develops will have very much in common with those constructed by authors who base their tests on one of the other hypotheses. In many respects, the processes of standardization will be the same; the same basic principles of constructing and testing will have to be observed. A variety of mental activities will have to be sampled; in the case of the multi-factor theory, in order to sample an adequate and representative number of the many minute factors; in the case of the group-factor theory, in order to sample the "primary" abilities and those "second-order" factors that might be found subsequently; in the case of the two-factor theory, in order to get an adequate sampling of the general factor.

The main practical differences arising from the theoretical differences will be found in the tests based on the group-factor theory, as compared with others. The differences will be these: the parts of the test based on group-factor theory must correspond with the factors or "primaries" and they must try to measure these factors in as "pure" a form as possible; the subtests in a scale based upon group-factor theory should have low intercorrelations; the test based on group-factor theory will emphasize the separate scores on each of the "primaries" and will provide a "mental profile," even though the group-factor test might also provide an over-all index. The Binet type of test, and other g-factor tests, on the other hand, consists of a variety of test materials which are a composite of abilities, yielding a mental

age and an intelligence quotient. Most group tests,³² while arranging their items according to type (sentence completion, arithmetical reasoning, word meaning, picture completion, form perception, etc.), are not organized on the basis of specifically defined factors; and, like the Binet type, they generally yield a single index of relative rank.

Since the several hypotheses regarding the nature of intelligence have thus far produced relatively few differences in practical test construction and application, the reader might well ask: "Why, then, be so concerned with definitions and theories, when the end-results are not radically different?" The answer to this question has several aspects. First, the student should be familiar with the thinking of psychologists in this field, as a background for his better understanding of the tests themselves. Second, it is through the interaction of the theoretical and applied that improvements and advances will be made. Third, it is possible that one or more of these theories will have increasing influence, in the future, upon test construction, testing practice, and test interpretation.³³

With these definitions and theories of intelligence in mind, we turn now to an examination and evaluation of some of the representative current tests of intelligence.

³² Excepting the "omnibus" type, in which items of various mental operations are placed in regular or irregular order, instead of being grouped in subtests, each containing items of a single kind.

³³ A useful presentation of some problems of test construction will be found in J. Loevinger, *A Systematic Approach to the Construction and Evaluation of Tests of Ability*, Psychological Monographs, Vol. 61, No. 4, 1947. See also L. L. Thurstone, "Psychological Implications of Factor Analysis," *The American Psychologist*, Vol. 3, 1948, pp. 402-408.

pacity. Although it has since been demonstrated that these measures have very little or no value for evaluation of the higher and more complex processes which we call "intelligence," Galton's work did nevertheless, affect the nature of test experimentation until about 1900, when the influence of Alfred Binet, the French psychologist was felt.

The kinds of psychological tests which were being used prior to 1900 are well represented by those with which J. McKeen Cattell was experimenting in the United States.¹ Ten of his most highly regarded tests at that time were the following:

- (1) Strength of grip, using the dynamometer
- (2) Rate of movement: the quickest time in which the hand can be moved through a distance of fifty centimeters
- (3) The smallest perceptible distance between two points on the skin, known as "two-point discrimination"
- (4) Amount of pressure necessary to cause pain by exerting pressure upon the forehead with a strip of hard rubber.
- (5) The smallest difference in weight which can be discerned, measured by requiring that two weights be lifted in succession
- (6) Speed with which an individual can react to a sound
- (7) Speed with which an individual can name ten specimens of four different colors arranged in haphazard order
- (8) The accuracy with which an individual can bisect a fifty-centimeter line
- (9) The accuracy with which an individual can reproduce an interval of ten seconds
- (10) Immediate rote memory, using a series of consonants

It is obvious that these are, for the most part, simple sensory and motor tests, with a little rote memory added. Other investigators in the United States and abroad were also experimenting with psychological tests, employing methods and materials similar to those of Galton and Cattell. Jastrow gave tests involving both touch and vision, tests of vision alone, tests of memory and tests of reaction time. Boas made physical measurements of children and also tested their vision, hearing, and memory. At the same time he obtained teachers' estimates of their pupils' intellectual acuteness. Gilbert used measures of height, weight, and lung capacity, tests of sensation, rapidity of tapping, reaction time, memory, and

¹ J. McK. Cattell, "Mental Tests and Measurements," *Mind*, Vol. 15, 1890, pp. 373-81.

THE BINET SCALES

HISTORICAL BACKGROUND

The development of mental tests was motivated by psychologists' interests in and attempts to measure individual differences in abilities. Experimental work in testing was begun in the 1860's by Francis Galton, the English geneticist and eugenicist, who wished to study the effects of heredity and environment upon the development of mental abilities, and thus the possibilities of "supplanting inefficient human stock by better strains." It was necessary, therefore, that he develop methods of testing psychological traits which would be more objective and valid than any means then available.

It was to be expected—in view of Galton's background as a biologist, the ascendancy of biological science during this period, and the emergence of experimental psychology as a separate discipline—that Galton should attempt to measure intelligence by means of the tests of imagery and sensory discrimination which he devised. For example, he used a test for the measurement of the delicacy of weight discrimination, and what is known as the "Galton whistle" for measuring susceptibility to high tones. Galton also suggested devices to be used for testing the other senses. In addition, he devised tests of imagery which were intended to discover the fidelity and detail with which a person could recall an earlier experience.

Apparently Galton assumed that the simpler measurable sensory capacities would show a significant correlation with intelligence, and that if these simple sensory measures were obtained they would afford a means of judging or predicting an individual's intellectual ca-

suggestibility. He also obtained teachers' ratings of the pupils' mentality.

Several other investigators during this period were beginning to introduce somewhat more complex materials and methods. Kraepelin and Oehrn, in Germany, used tests of perception involving the counting of letters, cancellation of letters, and detection of errors on the printed page; tests of memory involving digits and nonsense syllables; tests of association and of motor functions.

Muensterberg also devised tests of a more complex kind: reading aloud rapidly; giving rapidly the colors of named objects; rapidly naming and classifying animals, plants, and minerals; naming rapidly and classifying cloth, food, and parts of the body; naming rapidly ten simple designs and ten squares of colors; tests of addition; rapidly counting angles in irregular polygons; naming different odors. He also employed tests of memory for digits and letters after a single presentation; bisecting, judging, and reproducing lengths of lines; locating a sound; constructing a square and an equilateral triangle, with only the base of each given.

Some of Muensterberg's tests, it will be noted, were more complex and more varied than those of other experimenters; yet, like those of others, his were tests essentially of simple psychological processes, with a premium placed upon speed.

These relatively simple tests of sensory, motor, and memory capacities proved to be of very little value as measures to reveal intelligence. In the first place, their *intercorrelations* were very low, ranging generally from zero to only .20. And, in the second place, the results of these tests, when correlated with academic performance, yielded correlation coefficients of much the same magnitude, many of them being less than .10—hence useless for purposes of prediction. As a matter of fact, experimentation in the years that followed the early investigations has consistently confirmed the negligible or very low correlations found to exist between sensory and motor capacities, on the one hand, and the higher more complex functions, called "intelligence," on the other.

It is now generally recognized by psychologists that intelligence has little relationship to the elementary sensory and motor processes, and but a very moderate relationship indeed to capacity for rote memory (a correlation of about .30). Many infra-human animals have keen sensory discrimination. Mentally deficient children in the higher levels

of defect and children in the "borderline" group are not very inferior to normal children in respect to skin sensitivity, visual acuity, auditory acuity, reaction time, etc. Nor are intellectually gifted children superior to average in these respects. But in the capacities to learn, to organize and direct thinking, to adapt behavior, to comprehend problems and deal with abstractions, in levels of information possessed, in extent of curiosity about one's environment, these groups do differ very markedly indeed.

The reader should bear in mind these early kinds of tests, not only for historical purposes, but also in order to compare the early efforts with currently available tests and to be more clearly aware of the direction in which psychological testing has been moving.

THE EARLY WORK OF ALFRED BINET

The earlier work of Binet, the French psychologist, was along much the same lines as that of the American and German psychologists already mentioned. He used tests of tactual discrimination, reaction time, visual discrimination, auditory discrimination of time intervals, reproducing letters and numbers from memory, and so on. But though he experimented with these materials until about 1900, he had begun to doubt, some years earlier, the value of continuing with them.

Although some of the mental activities which Binet proposed to measure and with which he was experimenting were as yet vague, he did nevertheless point out the direction in which mental tests should and in fact did develop; that is, that the higher and more complex mental functions must be measured rather than the simpler sensory and motor processes.

Binet and his collaborators objected to the kinds of psychological tests which followed Galton's work, on the ground that they were too simple in character and would contribute little to the understanding of differences among persons in respect to the higher mental functions. Binet maintained, furthermore, that intelligence is expressed *not* in the form of simple, segmental responses, but rather as a combined mental operation wherein whatever processes are involved operate as a *unified whole*.

It is in these higher functions that individual differences are most marked; it is these which distinguish individuals most significantly and characteristically in daily activity, whereas it is in the simpler

sensory and motor processes that persons differ least significantly. Binet was quite ready to grant that the simpler processes, such as those already mentioned, lent themselves to more precise measurement and, therefore, yielded more constant results. Yet his interests were strongest in individuals rather than in sensory and motor processes as such. Thus, he was ready to sacrifice the greater quantitative precision of sensory-motor tests in order to obtain more valid evi-



FIG. 4.1. Alfred Binet (1857–1911).

dence of the integrated mentality of the individual; for he argued that in the measurement of the higher functions, the greatest precision of measurement, though desirable, was not as essential as in measuring the simpler processes, because of the very fact that individuals differ much more markedly in the former.

Binet emphasized that his proposed scale, testing the higher and more complex mental functions, would not measure in a physical sense, in the same way, for example, that the length of a line is measured. His tests would, however, yield "a classification, a hierarchy among diverse intelligences; and for the necessities of practice this classification is equivalent to a measure."²

Binet and his collaborators were interested in establishing the ex-

² A. Binet, *The Development of Intelligence in Children*, trans. by E. S. Kite, Vineland Training School, 1916, p. 40.

tent and nature of variations of mental functions from one individual to another, and in the determination of the interrelations of the several functions *within* the individual. In 1896, therefore, Binet and Henri (a collaborator) published their studies of the following functions: memory, the nature of mental images, imagination, attention, comprehension, suggestibility, esthetic appreciation, moral sentiments, muscular strength, strength of will, motor skill, and visual judgment. These, they believed, were functions ". . . which differ much from one individual to another, and are such that knowledge of their state for an individual gives us a general idea of this person, permits us to distinguish him from other individuals belonging to the same milieu."³ Here, then, we have the beginnings of the tests which a few years later proved so useful in the construction of the Binet scales and subsequently in the several revisions developed for use in the United States.

In 1904 a practical situation arose in which Binet had an opportunity to apply his principles with regard to differentiating individuals and to make an outstanding contribution to the study of mental abilities and individual differences. The French Minister of Public Instruction appointed a commission to recommend measures to be taken in the education of mentally subnormal children in the schools of Paris, for it was recognized that these children were unable to profit from the usual instruction. The plan was, therefore, to eliminate subnormal children from ordinary schools and to provide adapted instruction in a special school. Admission to the special school was to be based upon a medical and a pedagogical examination. Obviously, the first device needed was some objective means of selecting pupils of subnormal mentality. Subjective opinions were worse than useless; for not only was there absence of agreement among different so-called "experts," but serious injustices could result in numerous cases.

THE 1905 BINET-SIMON SCALE

It was to meet this problem that Binet constructed his first scale to evaluate children's intelligence levels. This instrument is known as the 1905 Binet-Simon Scale.⁴ In it we find a fundamental

³ For details, see A. Binet and V. Henri, "La Psychologie Individuelle," *L'Année Psychologique*, Vol. 2, 1896, pp. 411-465. For most of Binet's contributions, see *L'Année Psychologique*, Vols. 1-17.

⁴ Called a "scale" because the test items are arranged in order of increasing difficulty.

conception underlying all tests by means of which mental abilities of children are measured. *The principle is that we may identify differences in mentality, differences in degrees of brightness or dullness, with differences in levels of development as represented by the average capacities of children of various ages.* Thus, if we know the levels of intellectual performance of typical, or normal, children at each age, we can determine in the case of any individual child the extent to which his mental development is accelerated or retarded, or whether it is just about at the average level for his age at any given time.

In the 1905 scale, this conception was only crudely implemented; but in time it became more precise and has since taken the form of indexes already discussed in Chapter 2: namely, mental age, percentile ranks, decile ranks, standard scores, intelligence quotients, and others.

The thirty items, in order of increasing difficulty, comprising the 1905 scale follow: ⁵

(1) Visual coordination—degree of coordination of movement of head and eyes as a lighted match is slowly moved before subject's eyes

(2) Prehension provoked by tactual stimulus—a small cube of wood is placed on back or palm of the subject's hand to see if he grasps it and carries it to his mouth, and coordination of movements is to be noted

(3) Prehension provoked visually—cube of wood is placed within subject's reach by examiner who notes whether subject grasps it

(4) Recognition of food—a small piece of chocolate and a piece of wood of same dimensions are shown successively, and Signs of recognition of food and efforts to take it are noted

(5) Seeking food when slight mechanical difficulty is interposed—a small piece of chocolate, wrapped in a piece of paper, is given to the subject, and his manner of separating the food from the paper is noted

(6) Execution of simple directions and imitations of simple gestures

(7) Verbal knowledge of objects—parts of the body (head, ear, nose, etc.) are indicated by the subject, and common objects (key, string, cup) are handed to examiner on request

(8) Verbal knowledge of objects in a picture, as shown by pointing out objects, the names of which are given

⁵ For a detailed description, see G. M. Whipple, *Manual of Mental and Physical Tests*, Baltimore: Warwick and York, 1910, pp. 475 ff.

- (9) Naming of objects designated in a picture
- (10) Comparison of lengths of two straight lines, pointing out the longer
- (11) Repeating three digits immediately after hearing the series once
- (12) Comparison of weights; identical-appearing blocks of wood weighing 3 and 12 grams, 6 and 15 grams, 3 and 15 grams
- (13) Suggestibility—asking the subject for an object that is not present (modification of 7); asking subject to point to a nonexistent object in a picture, designated by a nonsensical word (modification of 8); comparison of lines of equal length (modification of 10)
- (14) Definitions of familiar objects: such as, house, horse, fork
- (15) Repetition of sentences having fifteen words each, after hearing each one only once
- (16) Giving the differences between two common objects: e.g., wood and glass, a fly and a butterfly
- (17) Immediate recall of pictures of familiar objects—pictures of thirteen common objects are shown for thirty seconds, after which the subject names as many as he can recall
- (18) Drawing from memory two different geometric designs which have been shown simultaneously for ten seconds
- (19) Repetition of series of digits, beginning with a series of three and proceeding until the subject's limit is reached
- (20) Giving resemblance between common objects: e.g., a wild poppy and blood; an ant, a fly, a butterfly, and a flea
- (21) Rapid comparison of lengths of lines: a line of 30 cm. is compared with fifteen others varying from 31 to 35 cm.; then a line of 100 cm. is compared with twelve others varying from 101 to 103 cm.
- (22) Discriminating and arranging in order five weights—3, 6, 9, 12, 15 grams—all being of equal size
- (23) Recall of weights—one of the weights in test 22 is removed, the remaining weights are scrambled, and the subject is asked to identify the missing weight or gap in the series
- (24) Giving rhymes to selected words
- (25) Sentence completion—supplying the correct word to complete a sentence
- (26) Devising a sentence to include three given words: e.g., Paris, gutter, fortune
- (27) Comprehending and giving replies to twenty-five problem-questions graded in difficulty: e.g., What is the thing to do when you are sleepy? Why is it better to continue with perseverance what one has started than to abandon it and start something else?
- (28) Reversing the hands of a clock, to be done from memory; e.g., giving the time it would be if the large and the small hands were interchanged at four minutes to three. The subjects who succeed were given the more difficult problem of explaining why the precise transposition indicated is impossible.

(29) Drawing lines to show the folds and cut-out of a piece of paper that has been quarto-folded and from which a triangular piece has been cut

(30) Giving definitions and distinctions between paired abstract terms: e.g., sad and bored

Although this set of tests was not separated into age groups, Binet did indicate several differentiating levels. Question number 6 was the upper limit of idiots (adult); question 9 was the upper limit of ordinary three-year-old children; number 14 was the limit of ordinary five-year-old children; number 16, that of imbeciles (adult); test 23, the most probable limit of morons (adult), although test 27 was regarded as having great value in revealing the moron. In addition, the authors reported a number of qualitative and quantitative differences in replies to many of the questions, thus distinguishing between 7- and 9-year levels, on the one hand, and 9- and 11-year levels, on the other.

The order of tests in the 1905 scale was experimentally determined, for it was established after being used with children in the primary schools and in an institution for the mentally deficient (the Salpêtrière). The children in primary school were regarded as "normal" on the basis of the fact that they were in grades just right for their ages—neither advanced nor retarded. Binet and Simon report that many such children were tested, but norms for the scale were based upon records of only ten cases in each of the following age-groups: 3, 5, 7, 9, and 11 years.

While admittedly rather crude and tentative, this first scale enabled Binet and Simon to classify idiots, imbeciles, and morons in a more objective manner than had been possible before.

Furthermore, in the foregoing list of thirty items, the reader will find many types which have since been developed, standardized, and included in a large number of current psychological tests, from those designed for babies to those intended for adult levels.

It is significant to note, also, that while Binet wanted to devise a scale that would yield age ratings, he was equally concerned with the *quality* of judgment and reasoning shown by the subject in the course of the examination. Binet was thus using the test situation as an opportunity for a clinical interview—a practice which is becoming increasingly widespread and of increasing importance in reports of psychological examinations by present-day clinical psychologists.

THE 1908 BINET-SIMON SCALE

Binet and Simon recognized the defects of the first scale. They recognized that an improved scale would have to provide more valid norms, based upon a larger and more representative sampling of children at each age; that tests for each age within the limits of the scale would have to be included to achieve finer units of measurement and greater accuracy. Their own subsequent investigations and those of other psychologists resulted in a new form of the test, known as the 1908 scale, in which the items are grouped at the appropriate age-levels, from 3 years to 13 years.⁶

Age 3

- (1) Points to nose, eyes, mouth
- (2) Repeats sentences of six syllables
- (3) Repeats two digits
- (4) Enumerates objects in a picture
- (5) Gives family name

Age 4

- (1) Knows own sex
- (2) Names certain familiar objects shown to him (key, knife, penny)
- (3) Repeats three digits
- (4) Perceives which is the longer of two lines 5 and 6 cm. in length

Age 5

- (1) Indicates the heavier of two cubes (3 and 12 grams; 6 and 15 grams)
- (2) Copies a square
- (3) Constructs a rectangle from two triangular pieces of cardboard, having a model to look at
- (4) Counts four coins
- (5) Repeats a sentence of ten syllables

Age 6

- (1) Knows right and left; indicated by showing right hand and left ear
- (2) Repeats sentence of sixteen syllables

⁶ A. Binet and T. Simon, "Le développement de l'intelligence chez les enfants," *L'Année Psychologique*, Vol. 14, 1908, pp. 1-94. See also the translation of Binet's publications by E. S. Kite, *The Development of Intelligence in Children*, Baltimore: Warwick and York, 1916.

- (3) Chooses the prettier in each of three pairs of faces (esthetic comparison)
- (4) Defines familiar objects in terms of use
- (5) Executes three commissions
- (6) Knows own age
- (7) Knows morning and afternoon

Age 7

- (1) Perceives what is missing in unfinished pictures
- (2) Knows number of fingers on each hand and on both hands without counting
- (3) Copies a written model ("The little Paul")
- (4) Copies a diamond
- (5) Describes presented pictures
- (6) Repeats five digits
- (7) Counts thirteen coins
- (8) Identifies by name four common coins

Age 8

- (1) Reads a passage and remembers two items
- (2) Adds up the value of five coins
- (3) Names four colors: red, yellow, blue, green
- (4) Counts backwards from 20 to 0
- (5) Writes short sentence from dictation
- (6) Gives differences between two objects

Age 9

- (1) Knows the date: day of week, day of month, month of year
- (2) Recites days of week
- (3) Makes change: four cents out of twenty in play-store transaction
- (4) Gives definitions which are superior to use; familiar objects are employed
- (5) Reads a passage and remembers six items
- (6) Arranges five equal-appearing cubes in order of weight

Age 10

- (1) Names the months of the year in correct order
 - (2) Recognizes and names nine coins
 - (3) Constructs a sentence in which three given words are used (Paris, fortune, gutter)
 - (4) Comprehends and answers easy questions
 - (5) Comprehends and answers difficult questions
- (Binet considered item 5 to be a transitional question between ages 10 and 11. Only about one half of the ten-year-olds got the majority of these correct.)

Age 11

- (1) Points out absurdities in statements
- (2) Constructs a sentence, including three given words (same as number 3 in age 10)
- (3) Gives any sixty words in three minutes
- (4) Defines abstract words (charity, justice, kindness)
- (5) Arranges scrambled words into a meaningful sentence

Age 12

- (1) Repeats seven digits
- (2) Gives three rhymes to a word (in one minute)
- (3) Repeats a sentence of twenty-six syllables
- (4) Answers problem questions
- (5) Interprets pictures (as contrasted with simple description)

Age 13

- (1) Draws the design made by cutting a triangular piece from the once-folded edge of a quarto-folded piece of paper
- (2) Rearranges in imagination the relationship of two reversed triangles and draws results
- (3) Gives differences between pair of abstract terms: pride and pretension.

There are several obvious differences between the 1905 scale and that of 1908. In the former, there are thirty test items; in the latter, fifty-nine. The latter does not include the first six items of the 1905 scale which are at the infant level; some other items of the 1905 scale have been eliminated, and many new ones have been added. As compared with the 1905 scale, the age-range extends higher in the 1908 scale. There are specific groups of items for each age (thus permitting a more accurate rating of individuals), and a greater variety of mental processes is tested.

In the 1908 scale, there are also two new and very significant contributions to the theory and practice of mental testing and test construction: (1) the tests, after experimentation, were standardized by being grouped into appropriate age-levels (Binet's method is explained below); (2) the concept of mental age is employed for the first time.⁷

The principal criterion employed by Binet and Simon in the standardization and age-placement of tests was this: in general, a test was

⁷ Although mental age is employed here for the first time, the concept itself had been arrived at by Binet in 1905.

placed at the year-level where it was passed satisfactorily by from two thirds to three fourths of a representative group of children of that age. The *ideal* standard was to place a test at a year-level where it was passed by *seventy-five* percent of that age-group. Binet's reason for setting this ideal criterion is a sound one and is made clear by reference to a symmetrical bell-shaped curve, which is approximated by most distributions of intelligence-test scores. The middle 50 percent of the group are most nearly alike, most nearly homogeneous in respect to the abilities being measured, as is obvious from the concentration of these 50 percent within a relatively narrow range, or variation, of scores. Otherwise stated, those individuals constituting the middle 50 percent of the distribution are the *typical* persons of the age group; hence, their test performance should be regarded as typical or normal for their age. If the middle 50 percent of a given age are able to pass a test, then that same test can be passed by the 25 percent who are above the middle group in ability, making a total of 75 who are able to pass the test.

In actual experience, however, it has been practically impossible to devise tests that will exactly satisfy this criterion of 75-percent passing. Fortunately, there are other criteria of validity which are also of primary significance, so that tests are retained if they approximate the 75-percent criterion, and if they demonstrate their value by also satisfying other demands, such as distinguishing between groups of individuals of known ability (mentally deficient, average, and superior), showing appreciable or significant differences between percentages passing at successive age levels, and correlating fairly well with scholastic achievement. These aspects of validation will be more fully discussed in the following chapter, in connection with revisions of the Binet scale.

Binet and Simon standardized their 1908 scale after individual examinations of 203 Paris school children between the ages of 3 and 13 years. While this number is small and would be regarded as inadequate in present-day test-standardization procedures, the fact is that these French pioneers did set a pattern of standardization which is being followed today, with the use of considerable statistical refinement. For in addition to having suggested the criteria already mentioned, they also, in effect, used the symmetrical bell-shaped curve as a criterion, though without offering precise numerical values. They stated, simply, that the number of children testing above

age (superior) should equal the number testing below age (inferior), and the number testing at age, or normal, should be greater than the number who rank as either superior or inferior.

The mental age, with the 1908 scale, was found this way: the subject was credited with the age level at which he passed all the tests. To this basic level (now called the "basal year") an additional year's credit was added for every five tests passed at higher levels. The total was the subject's mental age. No credits were given for a fraction of a year; but in the 1911 scale (see below) the calculation of mental age was modified so as to include fractional parts. The reader will note that this method of deriving mental age is essentially the same as that used with the American age-scale revisions.

In spite of its imperfect standardization, in the 1908 Binet-Simon Scale and in the publications concerned with it will be found many of the important concepts and practices which have been employed for about forty years in the construction and use of psychological tests.

THE 1911 REVISION OF THE BINET SCALE

The 1908 scale created considerable interest among psychologists in Belgium, Germany, England, Italy, Switzerland, and the United States. Their interest resulted in a number of valuable applications and evaluations of the Binet scale, accompanied by suggestions for revisions.

For the most part, criticisms and suggestions dealt with the age levels at which various items had been placed. It is not surprising that, in the first age-scale devised to measure intelligence, further and extensive applications and analysis of results should have revealed that a number of the items were misplaced. The principal criticism was that the tests at the lower age-levels were too easy, whereas those at the higher levels were too difficult, with the result that the former group were rated too high, while the latter were rated too low. In other words, standardization of the test had to be improved. Binet utilized the suggestions and criticisms of other psychologists, as well as the results of his own continued researches on the 1908 scale, the result being the 1911 revision.

Specifically, the major changes incorporated in the 1911 scale were the following: four of the tests at the 11-year level were raised to the 12-year level; all 12-year tests were raised to the 15-year level; the

three tests of year 13, plus two new ones, constituted the new adult level. Here and there, also, a few tests were placed in either a higher or a lower age-level. No tests were provided for the 11-, 13-, and 14-year levels.⁸ In addition to these changes, a fair number of tests found in the 1908 scale were omitted from the 1911 scale because they seemed to depend too much on school learning or on very incidental information.

At age-levels 3, 4, and 5 the tests are the same as in the 1908 version.

Age 6

- (1) Distinguishes between morning and afternoon
- (2) Defines names of familiar objects in terms of use
- (3) Copies a diamond
- (4) Counts thirteen *sous*
- (5) Distinguishes between pictures of ugly and pretty faces

Age 7

- (1) Shows right hand and left ear
- (2) Gives description of pictures
- (3) Executes three commissions given simultaneously
- (4) Gives value of 3 single- and 3 double-*sous*
- (5) Names four colors: red, green, yellow, blue

Age 8

- (1) Gives differences between two objects (from memory)
- (2) Counts backwards from 20 to 0
- (3) States omissions from unfinished pictures
- (4) Knows the date
- (5) Repeats five digits

Age 9

- (1) Makes change from 20 *sous*
- (2) Defines names of familiar objects in terms superior to use
- (3) Recognizes all nine [French] coins
- (4) Gives months of the year in correct order
- (5) Comprehends and answers easy problem-questions

⁸ Inasmuch as the rate of mental development appears to decrease appreciably after age 10, it becomes difficult to devise tests which will adequately distinguish between yearly levels. This difficulty was encountered also by the authors of the first Stanford Revision of the Binet-Simon Scale (1916); but in the second Stanford revision (1937) the authors were able to provide tests at yearly levels between ages 10 and 14. See the next chapter.

Age 10

- (1) Arranges five blocks in order of weight
- (2) Reproduces two geometric designs from memory
- (3) Criticizes absurd statements
- (4) Comprehends and answers difficult problem-questions
- (5) Uses three given words in two sentences

The method of scoring the 1911 scale was modified so that fractions of a year could be used in determining the mental age. Since there were five tests at each age level (except at age 4), each counted as two tenths of a year. Thus, if a child passed all tests at age 6, two at age 7, and one at age 8, his mental age would be 6.6 years.

According to Binet, a child whose mental age is equal to his chronological age is considered "regular" in intelligence; one whose mental age is higher is called "advanced"; and one whose mental age is lower is called "retarded." The degree of advancement or retardation in any instance is dependent upon the extent of the difference. Within about a year, however, William Stern was to suggest the use of the *intelligence quotient*, which has since been widely employed to indicate degree of acceleration or retardation in intelligence.

Binet's scales of 1908 and 1911 provided the stimulation and the basis for several adaptations and revisions in the United States. The authors of American revisions utilized Binet's principles and drew freely on his tests, as well as adding new ones and standardizing their instruments specifically for American children. These revisions will be presented in the following chapter.

SUMMARY

At this point we may very briefly summarize Binet's major contributions to the theory and practice of intelligence testing.

(1) If a psychologist is to develop a test of intelligence, he should first formulate a working conception and definition of intelligence, and then proceed experimentally. As a result of experimentation, new hypotheses will be developed; these in turn will influence later test construction; thus, both the conception and measurement of intelligence will undergo improvement and refinement. Binet's own conception of intelligence included mainly the following characteristics: ability to reason and judge well, to comprehend well, to take and maintain a definite direction of thought, to adapt thinking to the attainment of a desirable end, and to be autocritical.

(2) Intelligence must be measured by testing the higher, complex mental processes rather than relatively simple sensory and motor activities.

(3) Intelligence, being a complex, can be tested only by the use of a diversity of materials devised to evaluate the operations of mental processes as an integrated unit, rather than to measure the separate elements that might contribute to the complex functioning of intelligence. Though the Binet tests seem to be simple in conception and construction, they actually involve many complex mental activities; memory of several kinds, apperception, free association, orientation in time, language comprehension, ability with numbers, knowledge about common objects, constructive imagination, comparison of concepts, perception of contradictions, understanding of abstract terms, ability to meet novel situations, combining fragments into a meaningful whole.

(4) The tests included must be appropriate to the environment of those for whom they are intended.

(5) The tests were arranged in the form of a scale, from easiest to most difficult, and groups of tests were placed at appropriate age levels. The criterion was, ideally, that a test should be placed at a level where it was passed by three-fourths of that age group.

(6) The concept of mental age was introduced.

(7) The tests must be so standardized that the large middle group of average children (in the curve of distribution) will test "at age."

(8) Other criteria of validity were introduced: known groups, scholastic ratings, increase in percentage passing a test at successive age levels.

(9) The need of establishing the reliability of a test was recognized; Binet, therefore, made a few reliability studies with his 1911 scale.

Binet made not only these contributions; he indicated the very extensive uses to which psychological tests could be put in educational, social, vocational, and theoretical problems, for he regarded tests as tools for research and for scientific solution of important practical problems. Indeed, many of the researches and uses to which tests have since been applied are definitely along lines indicated by Binet, including, among others, the testing of prospective soldiers in order to eliminate the mentally unfit.

Binet did not regard his tests as final or as quite satisfactory; he did

not claim that they measured all aspects of personality; he emphasized that they must be supplemented by psychological and educational information derived by other means and from other sources. He did claim—and in this he has been supported by extensive subsequent use—that his test, and improved versions that should follow, can provide a very useful and reasonably reliable index of an individual's general intelligence, when the tests are administered and interpreted by qualified examiners.⁹

⁹ Alfred Binet died in 1911. His premature death deprived psychology of one of its great pioneers.

For a comprehensive study of Binet's psychology, see E. J. Varon, *Development of Alfred Binet's Psychology*, Psychological Monographs, Vol. 46, No. 3, 1935.

EARLY REVISIONS OF THE BINET-SIMON SCALE

FOUR EARLY REVISIONS

The two most widely known and used adaptations of the Binet scale in the United States are the Stanford revisions of 1916 and 1937. There were, however, four other revisions which, at one time or another, had some currency among psychologists but which today are infrequently used or are chiefly of historical interest. In 1908, H. H. Goddard published a translation of Binet's 1905 scale; and in 1911 he produced, for use in the United States, a revision of Binet's 1908 version. Yerkes published revisions in 1915 and 1923, in which the several types of items were grouped as subtests in a point scale (e.g., memory span for digits, analogies) instead of being placed at age-levels. Herring's revision appeared in 1922 and for some years was used as a valuable alternate in place of the 1916 Stanford scale. Kuhlmann's three revisions, 1912, 1922, and 1939, were extensive and elaborate in respect to standardization, scoring, and age-range covered. Thus it is clear that a significant amount of psychological work had been done prior to the publication of the 1916 Stanford scale and that several investigators continued their research and improvements on Binet's instrument for some years afterwards.¹

¹ Students interested in these historical aspects should consult the following: H. H. Goddard, "A Revision of the Binet Scale," *The Training School Bulletin* (Vineland, N. J.), Vol. 8, 1911, pp. 56-62; F. Kuhlmann, *A Handbook of Mental Tests*, Baltimore: Warwick and York, 1922, *Tests of Mental Develop-*

The full name of this test, The Stanford Revision of the Binet-Simon Intelligence Scale, is derived from the fact that the revision was made at Stanford University, under the direction of L. M. Terman. The construction of this scale was undertaken for the purpose of providing an instrument that would be adequately standardized and adapted for use in the United States. Its acceptance by psychologists and educators is attested by the fact that it was the most widely used individual scale until the revised Stanford-Binet appeared in 1937.

The 1916 scale includes 90 test items, covering an age range from 3 years to 14 years, with a group of test items added at the "average-adult" level and another at the "superior-adult" level. Of these 90 test items, 54 were adapted from the 1911 Binet scale, 5 from earlier Binet scales, 4 from other American tests, and 27 were new additions.

ment: *A Complete Scale for Individual Examination*, Minneapolis: Educational Test Bureau, 1939; R. M. Yerkes, et al., *A Point Scale for Measuring Mental Ability*, Baltimore: Warwick and York, 1915 and 1923; J. P. Herring, *Herring Revision of the Binet-Simon Tests*, Yonkers, N. Y.: World Book, 1923. For brief descriptions of these revisions see the 1950 edition of this textbook.

² L. M. Terman, *The Measurement of Intelligence*, Boston: Houghton Mifflin, 1916.

³ Terman, *op. cit.*, p. 53.

tient employed with this scale, an unselected group of children at each age should yield a median of 100.

Before the desired results were secured and this criterion satisfied, it was necessary to prepare three revisions of the scale. This involved the elimination of some test items, the shifting of others up or down in age level, and changes in scoring standards.⁴ "As finally revised," Terman states, "the scale gives a median intelligence quotient closely approximating 100 for the unselected children of each age from 4 to 14."

The test items above the age of 14 were based on examinations of 30 businessmen, 150 migrating unemployed men, 150 adolescent delinquents and 50 high-school students. These groups are not a representative cross-section of persons above fourteen years of age in the general population. This fact will help make clear why the 1916 scale was found to be unsatisfactory for use with older adolescents and with adults.⁵ The unsatisfactory quality of the scale at the upper ages was due also to inadequate sampling of abilities.

TABLE 14
Percents Passing Tests Located in Year VI:
1916 Revision⁶

Test	Ages				
	4	5	6	7	8
Right and left	40	50	71	86	95
Mutilated pictures	27	50	65	87	96
Counting 13 pennies	30	46	76	93	86
Comprehension	25	55	70	86	93
Naming four coins	25	47	74	91	95
Repeating 16-18 syllables	34	56	69	90	95

⁴ A theoretical or ideal percentage of passes for the placement of a test of a given age level was not used. Terman states, "We had already become convinced . . . that no satisfactory revision of the Binet scale was possible on any theoretical considerations as to the percentage of passes which an individual test ought to show in a given year to be considered standard for that year." (*Op. cit.*, p. 54.) Accordingly, a "trial-and-success" method was used in order to get the desired median mental age and IQ at each chronological age level. The same practice was followed in standardizing the 1937 revision.

⁵ In a personal communication Dr. Terman states that there were three tentative versions of the scale before the final one was published. The businessmen and high-school students were used in making the first tentative placement of tests at average and superior adult levels. The other adult groups were then used in subsequent rearrangement of test items.

⁶ From L. M. Terman and others, *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence*, Educational Psychology Monographs, No. 18, 1917, pp. 167-168. (By permission.)

In addition to the criterion of a significant increase in the percent passing a test item at successive ages, the following criteria were used in establishing validity of the 1916 scale.

First, in each age group, all the subjects tested were divided into the following three classes: those testing below 90 IQ, those testing between 90 and 109, and those testing 110 or above. Each test item was then examined to determine whether it was passed by a "decidedly higher" percentage of individuals in the superior IQ group than in the inferior. (The term "decidedly higher" was not defined by Terman.) Only those test-items which satisfied this criterion were retained. The following data are illustrative.

TABLE 15
Percents Passing Certain Tests
Chronological Age Constant ⁷

		IQ's		
		Below 96	96- 105	Above 105
Age 6	Counting thirteen pennies	40	77	96
Age 7	Describing pictures	48	52	80
Age 8	Giving similarities	44	57	83
Age 9	Making change	39	60	73
Age 10	Comprehension of problem situations	25	64	76

Second, after the scale had been developed, the IQ's obtained with 504 school children were compared with their scholastic ratings, as graded by their teachers, on a five-point scale; namely, very inferior, inferior, average, superior, very superior. Moderate agreement was found between intelligence quotients and school ratings, the coefficient of correlation being .48—close enough so that Terman and his colleagues concluded there was no justifiable "serious suspicion as to the accuracy of the intelligence scale."

Third, the relation between IQ and grade progress was studied for the children on whom the scale was standardized. A "fairly high" correlation was found, but there were also some "astounding disagreements," inasmuch as a given mental age level was found in a wide range of grades. For example, a mental age of nine was found in all grades from 1 to 7. Terman states, however, "When the data were examined, it was found that practically every child whose grade failed

⁷ From Terman and others, *ibid.*, p. 133. (By permission.)

to correspond fairly closely with his mental age was either exceptionally bright or exceptionally dull. Those who tested between 96 and 105 IQ [the average children] were never seriously misplaced in schools." ⁸

Reliability. Following its publication, the Stanford-Binet was subjected to numerous studies in order to determine its reliability by the method of self-correlation. The correlation coefficients, which in such studies will vary with the size and constitution of the experimental group, ranged from about .80 to .95. Such coefficients are regarded as highly satisfactory indexes of reliability.

The 1916 Scale. The reader will have noted that no essentially new concepts or principles have been added in the 1916 Stanford-Binet scale, as compared with Binet's own. Terman and his colleagues did, however, extend, refine, and adapt the Binet scales, so that the 1916 revision was a better standardized, hence more valid and reliable, instrument.

The complete list of tests of the 1916 Stanford-Binet follows.⁹ Throughout the scale, those items designated "Al.," instead of being numbered, are alternates, to be used in place of one of the numbered items, where the examiner, for any reason, believes a numbered item to be inappropriate.

Age 3

- (1) Points to parts of body
- (2) Names familiar objects
- (3) Enumerates objects in pictures
- (4) Gives sex
- (5) Gives last name
- (6) Repeats six to seven syllables
- (Al.) Repeats three digits

Age 4

- (1) Compares lengths of lines
- (2) Discriminates between geometric forms
- (3) Counts four pennies
- (4) Copies a square
- (5) Comprehends and solves problem situations
- (6) Repeats four digits
- (Al.) Repeats twelve to thirteen syllables

⁸ Terman, *The Measurement of Intelligence*, p. 74.

⁹ Reproduced by permission of Houghton Mifflin.

Age 5

- (1) Compares weights
- (2) Names familiar colors
- (3) Makes esthetic comparisons of paired drawings of faces
- (4) Defines common words: use or better
- (5) Puts together a divided triangle
- (6) Carries out three commissions
- (Al.) Gives own age

Age 6

- (1) Knows right from left
- (2) Perceives missing parts in pictures
- (3) Counts thirteen pennies
- (4) Comprehends and solves problem situations
- (5) Identifies coins
- (6) Repeats sixteen to eighteen syllables
- (Al.) Knows morning from afternoon

Age 7

- (1) Knows number of fingers on each and both hands
- (2) Describes pictures
- (3) Repeats five digits
- (4) Ties a bow-knot
- (5) Gives differences between paired objects
- (6) Copies a diamond
- (Al. 1) Names days of week in correct order
- (Al. 2) Repeats three digits backwards

Age 8

- (1) Traces path to be followed in a systematic search for a lost object in a field
- (2) Counts backwards from 20 to 1
- (3) Comprehends and solves problem situations
- (4) Gives similarities between two things
- (5) Defines names of objects in terms superior to use
- (6) Defines twenty words from a vocabulary list
- (Al. 1) Identifies six coins
- (Al. 2) Writes short sentence from dictation

Age 9

- (1) Gives date: day of week, month, day of month, year
- (2) Discriminates between weights: 3, 6, 9, 12, 15 grams
- (3) Makes change in small amounts
- (4) Repeats four digits backwards
- (5) Makes up a sentence including three given words
- (6) Gives rhymes to three words

Early Revisions of the Binet-Simon Scale

- (Al. 1) Names the months of the year
- (Al. 2) Gives total value of a group of one-cent and two-cent postage stamps

Age 10

- (1) Defines thirty words from vocabulary list
- (2) Detects absurdities in statements
- (3) Reproduces two designs from memory
- (4) Reads a short passage and reproduces content
- (5) Comprehends and solves problem situations
- (6) Names any sixty words by free association
- (Al. 1) Repeats six digits
- (Al. 2) Repeats twenty to twenty-two syllables
- (Al. 3) Fits rectangular blocks into form-board

Age 12

- (1) Defines forty words from vocabulary list
- (2) Defines abstract words
- (3) Traces a path in systematic search (same problem as in year 8, but a superior plan is required here)
- (4) Rearranges dissected sentences into meaningful sentences
- (5) Interprets fables
- (6) Repeats five digits backwards
- (7) Interprets pictures
- (8) Gives similarities between three things

Age 14

- (1) Defines fifty words from vocabulary list
- (2) Discovers a rule in a paper-folding test (induction test)
- (3) Gives differences between a president and a king
- (4) Integrates given facts and arrives at a conclusion concerning them
- (5) Solves arithmetical reasoning problems
- (6) Reverses hands of clock, in imagination, and gives the hour
- (Al.) Repeats seven digits

Average Adult

- (1) Defines sixty-five words from vocabulary list
- (2) Interprets fables
- (3) Gives differences between abstract words
- (4) Solves problem of number of enclosed boxes (boxes within boxes) when shown only the large outside box
- (5) Repeats six digits backwards
- (6) Perceives the pattern of a code and uses it
- (Al. 1) Repeats 28 syllables
- (Al. 2) Comprehends problems involving physical relations

" Superior Adult

- (1) Defines seventy-five words from vocabulary list
- (2) Visualizes, imaginally, and draws appearance of a folded and cut piece of paper
- (3) Repeats eight digits
- (4) Repeats thought of a passage heard
- (5) Repeats seven digits backwards
- (6) Solves problems involving "ingenuity"

* **The Scoring Method.** Each age level from 3 years through 10, it will be noted, has six test items (plus the alternate which may replace one of the six). Each of these carries credit of two "months," so that the tests in each of the age levels provide a year's increment in mental age.

There are no tests at the eleven-year level, the reason being that the authors of the scale were, apparently, unable to devise tests that would indicate a one-year difference at this stage of mental development. This gap in the scale, it is believed, is due to the slowing down of mental development, thus decreasing the annual increments and making it more difficult to measure those increments by means of the then-available test items. Since the eight test items at age twelve cover a span of two years, each one carries a credit of three months in order to yield an average mental age of twelve, when added to the ten-year level.

The same explanation applies to the six test items at age 14, each of which gives a credit of four months toward mental age score.¹⁰

The tests and credits at the "average adult" level were devised so as to provide a median value mental age of 16. Yet each of the six tests at average adult level carries a credit of five months, so that a person who passed all of them would get a mental age of 16.5 years. In his volume describing the standardization of the 1916 revision, in explaining the limit for "average adult," Terman states that his data on mental ages of 62 adults, including 30 businessmen and 32 high school pupils, who were over 16 years of age, show ". . . that the middle section of the graph [of the distribution] represents the 'mental ages' falling between 15 and 17. This is the range we have designated as the 'average adult' level."¹¹

Those persons having mental ages above seventeen were designated

¹⁰ The 1937 revision provides a group of tests at age 11 and another at 13.

¹¹ Terman, *The Measurement of Intelligence*, p. 55.

The standard deviation (S.D.) of this distribution is about 12 points. (Compare this with the S.D. of 16 points of the 1937 revision.) This distribution, being a fairly symmetrical one, showed that the scale did differentiate between the several levels of mental capacity of the persons examined; at least, so far as concerns the mental processes being tested. It therefore strengthened the belief of Terman and many others that the 1916 Stanford-Binet had considerable validity. Another method used to represent the frequency with which different degrees of intelligence occur was to indicate the percentage of subjects at and above or at and below a given IQ, thus:

TABLE 17

Percentage Distribution of IQ's:
Stanford-Binet Scale, 1916¹³[illegible]

Although the foregoing percentages above or below certain IQ levels are not fixed or identical for all tests (e.g., the distribution for the 1937 Stanford-Binet is not identical with this), a table such as this is significant in that it provides one means of determining an individual's relative status in respect to the psychological processes being

¹³ From Terman, *The Measurement of Intelligence*, Boston: Houghton Mifflin, 1916, p. 78. (By permission.)

as "superior adults," the possible maximum mental age on this scale being 19.5. (Six tests, six months' credit each, added to the maximum of 16.5 attainable at the "average adult" level.) The method of scoring the 1916 scale is this: The examiner goes *down* in the test until the level is reached where the subject passes all items. This is called the "basal year." The examiner then proceeds *upward* in the scale until the level is reached where the subject fails all items. This level is called the "terminal year." As already stated, each test item carries specified credit, in terms of months, contributing to the mental age score. These credits are added to the age value of the basal year; the total is the mental age. For example: assume that in a given instance the basal year is 6; three test items are then passed at the 7-year level, giving additional credit of 6 months; two are passed at the 8-year level, giving further credit of 4 months; all are failed at the 9-year level. Thus, this subject's mental age is 6 years, 10 months.

Distribution of IQ's. A second score obtained with the 1916 Stanford-Binet is the intelligence quotient, the calculation and general nature of which have already been explained. Terman not only found the IQ for each individual examined, but he analyzed the distribution of intelligence quotients obtained by the persons on whom the scale was standardized.

Taking those subjects between the ages of 5 and 14 years, the distribution was found to be the following:

TABLE 16
Distribution of IQ's of 905 Unselected Children,
Ages 5-14 Years¹²

IQ	Percent of total
56-65	0.33
66-75	2.3
76-85	8.6
86-95	20.1
96-105	33.9
106-115	23.1
116-125	9.0
126-135	2.3
136-145	0.55

¹² From Terman and others, *op. cit.*, p. 40. (By permission.)

measured; for, as already explained, the IQ is an index having educational and clinical connotations.

On the basis of the distribution of intelligence quotients obtained with the 1916 revision, Terman also suggested the following classification:

TABLE 18
Suggested Classification of IQ's:
*Stanford-Binet Scale, 1916*¹⁴

IQ	Classification
Above 140	"Near" genius or genius
120-140	Very superior intelligence
110-119	Superior intelligence
90-109	Normal, or average, intelligence
80-89	Dullness
70-79	Borderline deficiency
Below 70	Definite feeble-mindedness

This classification is reproduced because it has been widely used and because the reader should be familiar with its source. Unfortunately, however, such classifications or labels have frequently been used uncritically, the erroneous assumption having been that there was some quality inherent in the classification itself that warranted the designation of "genius," or "superiority," or "dullness," etc. It has already been pointed out, for example, that not all persons of very high IQ's have original, creative mentalities; yet these are among the traits of the "genius." The particular IQ intervals and the names attached to them result from the judgment of the specialist making the classification. Some classifiers might choose to place the lower limit of "near genius or genius" at 150 IQ, or the upper limit of "feeble-mindedness" at 50 or 60. In short, tables of IQ classifications are essentially statistical.

Regardless of size of intervals or of their names, a classification is useful chiefly as a convenient device for purposes of research and analysis of data. It should not be used merely to label and pigeonhole an individual who has been examined and for whom an intelligence quotient has been obtained. The trend is away from stating test results in terms of MA and IQ alone; the trend is in the direction of evaluation of individual performances.

¹⁴ From Terman, *op. cit.*, p. 79. (By permission.)

Adult Mental Age and Adult Intelligence Quotient. Since the 1916 Stanford-Binet includes tests at the levels of average adult and superior adult, it was necessary to make provisions for the calculation of adult mental ages and intelligence quotients. These, however, present special problems.

We have already quoted Terman's reason for locating average adult performance in the mental age range of 15 to 17, with the assumed midpoint at 16 years. If this is correct, it means that the test performance of the average adult is equal to that of the average 16-year-old individual. Otherwise stated, it means that in the case of an *average* adult, his maximum level of measured intelligence is reached at the age of 16 and that there are no increments thereafter. Terman states that ". . . in so far as it can be measured by tests now available, [intelligence] appears to improve but little after the age of 15 or 16. . . . Although this point [at which intelligence attains its final development] is not exactly known, it will be sufficiently accurate for our purposes to assume its location at 16 years."¹⁵ Thus, until the process of decline sets in, *the average adult continues to have a mental age of 16*, according to the 1916 Stanford-Binet.

On the basis of this assumption, then, in the calculation of an IQ for a person who is 16 years of age, or older, the denominator in the formula ($IQ = \frac{MA}{CA}$) is always 16. Otherwise, if his actual CA were used, he would *appear* to be getting rapidly less and less intelligent with the succeeding years. For example, an average individual at the age of 16 will have an IQ of 100 (16/16). At the age of 18, he should still have an IQ of 100 even though, according to the tests being used, there has been no further measurable development of mental capacity; for the formula will still be 16/16. Now if, in the case of this same individual, we had continued using his actual CA as the denominator, his IQ at age 18 would be shown as about 89 (16/18); at age 20, it would be shown as 80 (16/20), and so on, while as a matter of fact there would ordinarily have been no such decline. Thus, by using the denominator of 16 in the IQ formula for all persons above age 16, if a person of 20 years and one of 60 years have the mental age of 16 years each, then each will be given an IQ of 100.

The reader has also noted that it is possible to attain mental ages

¹⁵ *Op. cit.*, p. 140.

above 16 on the 1916 revision, the maximum being 19.5 years at the level of superior adult. If the definition of mental age is borne in mind, it will be apparent at once that a mental-age rating which is higher than that of the *average* adult has a new and specialized meaning. It cannot have the same meaning as the term, mental age, does ordinarily. A mental age is defined as the level of mental development of the *average* or *typical group* of persons at that same chronological age. Thus, an MA of 10 represents the test performance and mental level of a group of average children of chronological age 10. Hence, if we assume that *average* or typical adults reach a mental age of 16, then to speak of a "mental age" above 16 is to introduce a new concept; for these latter "mental ages" are not derived from the performance or norms of average or typical persons. They are theoretical and hypothetical indexes devised to enable us to indicate higher than average mental levels and higher than average intelligence quotients.¹⁶ Thus, when a higher-than-average adult "mental age" is used, it is essential that the user be aware of the fact that a new and different concept is being employed.

The fact that the highest possible "mental age" that can be attained on this test is 19.5 years means that the highest IQ an adult can get is about 122 (19.5/16). This maximum reveals a serious inadequacy of the 1916 revision at the higher levels. What, for example, happens to the IQ of the 10-year-old child who has a mental age of 15, and an IQ of 150 (15/10)? Obviously, to maintain that IQ of 150 at age 16 or older, he must be able to attain a mental age of 24 (24/16); yet the scale permits a maximum MA of only 19.5, with an IQ of 122. The same would be true for this subject after age 16.

Criticisms of the 1916 Stanford-Binet. Experience with the 1916 Stanford-Binet demonstrated that it was inadequate as a measure of adult mental capacity. In fact, experience showed that its usefulness was restricted to ages between five and fourteen years, the range between five and ten years being the most satisfactory.

¹⁶ In their volume describing the 1916 revision (*op. cit.*) Terman and his collaborators do not report how they arrived at their mental-age levels above those of average adult. In their 1937 revision this has been done by extrapolation and by providing for a distribution of adult IQ's which should correspond to distributions for pre-adult levels.

This revision was also criticized on several other grounds. First, since the scale was finally standardized on the basis of results obtained with approximately 1000 native-born children in California, its use with *all* groups of children in *all* parts of the United States seemed to many educators and psychologists to be a practice of doubtful validity; for it was held that the 1000 California subjects were not necessarily representative of the child population of this country. There is merit in the criticism, yet it must be recognized that this scale proved to be very useful in many parts of the United States, when employed and interpreted by examiners who were familiar with its assumptions and construction, and who, at the same time, were familiar with the backgrounds of the subjects they were examining.

Second, the scale was criticized as being much too heavily weighted with verbal and abstract materials, thus penalizing the individual who, for whatever reason, had been handicapped in developing his "verbal intelligence" through the medium of the English language. Terman's reply to this criticism was that intelligence at the verbal and abstract levels is the highest form, the *sine qua non*, of mental ability. Indeed, he defined intelligence as the ability to deal with abstract terms, and to do conceptual thinking.

This criticism of the scale was warranted, nevertheless; for children who are handicapped by lack of opportunity to acquire and develop the use of the English language are at a serious disadvantage and get spuriously low ratings on psychological tests which emphasize "verbal intelligence." Such children would include: (1) those who have developed in homes where only a foreign language is spoken; (2) those who are handicapped by serious visual or auditory defects; (3) those handicapped by sensory anomalies (reversals, inversions, mirror-writing, poor sound discrimination) which seriously interfere with their learning to read; (4) those who are too young—say, below age four or five—to be tested adequately by means of verbal materials almost exclusively.

Third, the 1916 scale was found to be defective at some points in respect to procedures in administering and scoring, thus detracting from its objectivity and from the comparability of results obtained by different examiners.

In view of these criticisms, it was to be expected that other scales should be developed, especially those of the "performance" and "non-

verbal" type, which would obviate or minimize the second criticism. These will be presented and discussed in a later chapter.

It was to be expected, also, that the 1916 Stanford-Binet itself should undergo revision in the light of experience, criticism, and accumulated data. Such a revision, begun about ten years after the original Stanford-Binet appeared, was published in 1937.

THE 1937 REVISION OF THE STANFORD-BINET SCALE

DESCRIPTION OF THE 1937 SCALE

This scale differs from that of 1916 in many details, but it does not differ in its essential and basic conceptions.¹ As the authors themselves state, "The revision utilizes the assumptions, methods, and principles of the age scale as conceived by Binet." The authors, however, do regard it as a better-standardized and more useful scale than its predecessors. The principal differences and modifications follow.

The 1937 scale has two equivalent forms (L and M), each of which contains 129 test items, as compared with the 90 items in the first Stanford-Binet. Items that proved unsatisfactory in the original were eliminated, and new ones were added.

The 1937 scale extends downward to the level of age 2, and upward through three levels of "superior adult" (known as Superior Adult I, II, and III), thus increasing its usefulness.

The levels below age 5 and those above age 14 have been more carefully and validly standardized.

Scoring standards and instructions for administering the tests were improved.

¹ L. M. Terman and M. A. Merrill, *Measuring Intelligence*, Boston: Houghton Mifflin, 1937. See also R. Pintner and others, *Supplementary Guide for the Revised Stanford-Binet Scale (Form L)*, Applied Psychology Monographs, No. 3, 1944. This monograph presents, for purposes of comparison, a collection of responses to test items but does not duplicate those provided in the Terman-Merrill manual.

From the age of 2 to age 5, this scale provides groups of test items at half-year intervals. Thus more accurate and more highly differentiating test results are obtainable. The half-yearly intervals are possible because the rate of mental growth is most rapid in the earlier years and, therefore, the more rapid periodic increments are susceptible to testing.

Groups of tests are provided at ages 11 and 13, whereas there were none at these levels in the 1916 scale for reasons already stated.

Although the 1937 scale, like that of 1916, is predominantly verbal in character, it does provide more performance and other nonverbal materials at the earlier age levels, especially through the age of four years. Performance materials are those with which the subject has to *do* something; for example, building a pattern or making a design with blocks, or filling in a form board with the variously shaped blocks. Other nonverbal materials include such activities as copying a geometric figure, completing the picture of a man, discriminating between forms, etc. In all of these, of course, verbal ability is a factor to the extent that verbal directions must be understood. In these tests, verbal ability can also operate as a factor if the subject is familiar with the names of the objects or geometric figures and is thus facilitated in his manipulation or classification of them.

The 1937 scale has been standardized on a more carefully chosen and much more extensive group of subjects. The base of the standardization population was broadened, and its component members are regarded as more representative of the population.² *But only American-born white subjects were used in the standardization of this scale*, the total number being approximately 3000. The subjects were chosen from eleven states in several widely separated areas of the country, and an effort was made to have the subjects from homes which, occupationally and socially, would be representative of the population at large in the United States.

VALIDATION

The test items were chosen on the basis of their validity, ease and objectivity of scoring, economy of time in administering, interest to the subjects, and need for variation in types of materials.

² Up to the age of 5, the number of subjects used was 100 to each half-year level; from 6 to 14 years of age, 200 at each year; from 15 to 18 years, 100 at each year.

Of the foregoing, *validity* is of primary significance. In this revision, a criterion of basic importance in judging validity of test items was increase in percentage of successful performance with increasing age. This criterion was applied in two ways: first, by requiring an appreciable increase in the percent passing a given item in successive ages (as in the 1916 scale); and second, by finding "a weight based on the ratio of the difference to the standard error of the difference between the mean age (or mental age) of subjects passing the test and of subjects failing it."³ Stripped of its statistical terminology, what this quotation means is that the difference between the average age (chronological or mental) of subjects passing an item, on the one hand, and the average age of subjects failing that item, on the other hand, must be statistically significant. This is essentially an "age criterion." In this connection see Table 19.

TABLE 19
Percent Passing Test Items Located in Year VI,⁴
Form L

Item	Ages						
	4	4½	5	5½	6	7	8
1.	3	15	36	50	67	89	97
2.	11	29	44	55	70	86	95
3.	11	26	46	53	69	86	96
4.	3	11	43	48	71	94	96
5.	16	29	47	51	73	94	95
6.	26	44	52	61	81	91	93

A second criterion of major importance in the retention of an item was its correlation with the *total scores* of the individuals of the age level at which the test item is located. Table 20 presents the distribution of correlation coefficients (biserial) for both forms L and M.

The calculated median of the coefficients for form L is approximately .69; the middle 50 percent of the coefficients fall between approximately .51 and .73. The range for the whole set of coefficients is from .28 (memory for designs, year 11) to .89 (abstract words, year 11; and vocabulary, year 14).

³ Terman and Merrill, *op. cit.*, p. 9. See also V. V. Fleming, "A Study of the Subtests in the Revised Stanford-Binet Scales, Forms L and M," *Journal of Genetic Psychology*, Vol. 64, 1944, pp. 3-36.

⁴ From Q. McNemar, *The Revision of the Stanford Scale*, Boston: Houghton Mifflin, 1942, p. 92. (By permission.)

The highest coefficients were found for IQ's below 70 (.98), the lowest for IQ's above 130 (.90), and intermediate coefficients for IQ's near 100 (.92). Age levels above six years showed greater reliability (.93) than those below six (.88), calculated for separate age-groups.

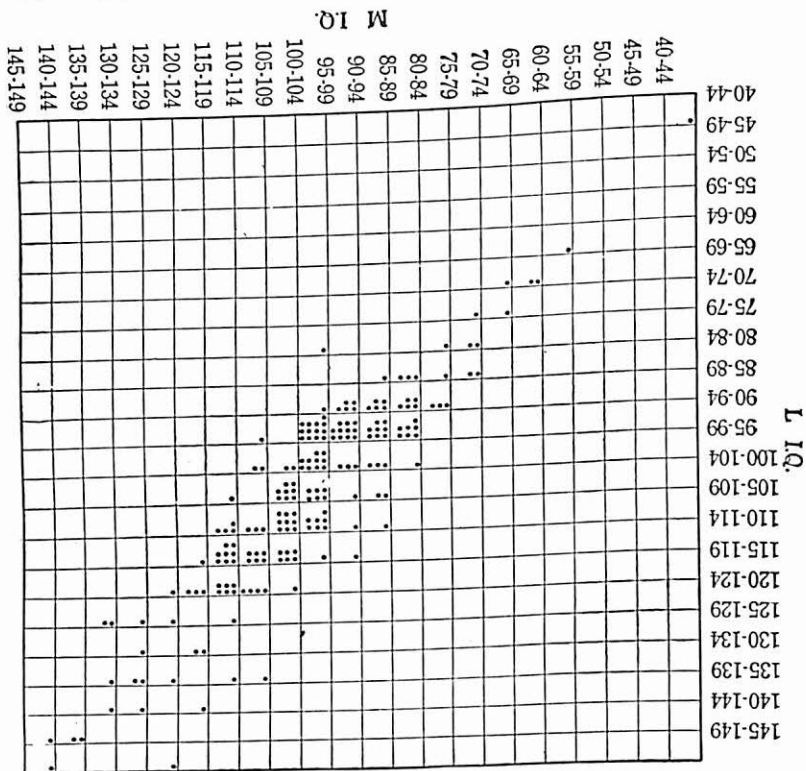


FIG. 6.1. Scatter Chart of Correlation between Form L and Form M IQ's at Chronological Age 7. $r = .91$. From Terman and Merrill, op. cit., p. 45. (By permission.)

These coefficients are of the same general order as those found by others in subsequent studies. (See Figure 6.1).

DETERMINING MENTAL AGE AND INTELLIGENCE QUOTIENT

The scoring method is exactly the same in principle as that used with the 1916 scale for the determination of mental age and intelligence quotient. There are, however, a few differences in the details.

On from M, the calculated median coefficient is approximately .64; the middle 50 percent of the coefficients fall between approximately .51 and .71. The range for this whole set of coefficients is from .27 (memory for stories, year 13) to .91 (abstract words, year 13). Of the 258 coefficients, 201, or very nearly 78 percent, are .50 or higher. This fact and the data presented in a later section in this chapter (*Analysis of Functions Tested*) provide strong evidence that the Stanford-Binet Scale measures "general ability" by means of test items that have psychological functions in common to a high degree.

TABLE 20

Distribution of Correlation Coefficients (Biserial)
for Each Item with Total Scores^a

r	Frequency	
	Form L	Form M
.20-	1	2
.30-	5	7
.40-	21	21
.50-	28	23
.60-	32	42
.70-	34	23
.80-	8	9
.90-	129	129

After selection of the tests that were to be used, one other empirical procedure was employed in locating each test item at an appropriate age level. The items were rearranged until it was found that they would yield for each group of subjects a mean mental age that was identical with their mean chronological age, giving a mean IQ as close as possible to 100. Six successive revisions were necessary to achieve this for Form L. Then, "... it was possible to achieve at once an equally good result with Form M by arranging its tests so as to match those of Form L at each age level with respect to difficulty, validity, and shape of curves of percents passing by age."⁶

RELIABILITY

Comparing IQ's obtained with Forms L and M, Terman and Merrill report reliability coefficients of correlation from .90 to .98.

^a Based upon data in Q. McNemar, *op. cit.*, Tables 53 and 54.

⁶ Terman and Merrill, *op. cit.*, p. 23.

Whereas the maximum mental age attainable on the 1916 Stanford-Binet was 19 years and 6 months, the maximum on the 1937 revision is 22 years and 10 months. It will be recalled that with the first Stanford-Binet scale, a maximum CA of 16 was used in the denominator to determine the IQ of an individual of 16 years of age or older. In the 1937 scale, the maximum CA in the denominator is 15. Thus, the highest possible IQ attainable by a subject who is 15 years of age or older is $152 (22^{-10}/15)$.

In the superior-adult levels, the test items were selected and their credits allotted (in terms of months) in such a manner as to make the IQ distribution of ". . . the older subjects resemble closely those of the younger, as presumably should be the case on an ideal scale."⁷

TABLE 21

*Correction of CA Divisor: 1937 Stanford-Binet Scale*⁸

Actual CA	Corrected CA Divisor
13-0	13-0
13-3	13-2
14-0	13-8
14-6	14-0
15-0	14-4
15-6	14-8
16-0	15-0

In order to achieve this desired goal, it was necessary for the authors to make adjustments in the denominator of the IQ formula, beginning at the age of 13 years and 2 months. The reason given for this adjustment is that it is extremely difficult, perhaps impossible, to escape the effects of selection of subjects at the upper ages in standardizing a scale. The selection generally is such as to include the average range and the higher mental levels but not the lowest, since the less intelligent individuals tend to leave school earlier than the others. Hence, the *norms* of test-performance of the older groups, it is argued, tend to be higher than they should be for an unselected sampling. These higher norms, in turn, tend to reduce the intelligence quotients of subjects in the older groups. It is this effect that the authors of the scale sought to correct by means of adjusting the denominator.

⁷ *Ibid.*, p. 30.

⁸ From Terman and Merrill, *op. cit.*, p. 31. (By permission.)

While Terman and Merrill believe they minimized the effects of selection, they were not wholly eliminated. Therefore, after a "trial-and-success" process directed toward making IQ distributions of older age groups resemble closely those of younger groups, the procedure adopted was this: from age 13 to age 16, the cumulative dropping of one out of every three additional months of chronological age, and all of it after 16. In substance, this practice is tantamount to saying that

TABLE 22
IQ Means Adjusted for 1930 Census Frequencies
of Occupational Groupings^a

Composite of Forms L and M, Stanford-Binet			
Age	N	Raw	Smoothed
2	76	102.1	
2½	74	104.7	103.3
3	81	103.2	104.1
3½	77	104.3	102.2
4	83	99.2	101.6
4½	79	101.2	100.8
5	90	101.9	100.4
5½	110	98.2	100.0
6	203	100.0	99.8
6½	202	101.2	100.8
7	203	101.1	102.0
7½	204	103.6	102.7
8	201	103.5	103.0
8½	204	101.9	102.2
9	202	101.2	101.6
9½	204	101.8	101.0
10	202	100.0	101.3
10½	107	102.0	101.3
11	102	101.8	103.3
11½	109	103.2	103.8
12	101	106.3	

average adult mental age, on the 1937 scale, is 15. Table 21 gives a few examples.

When, therefore, this scale is used with subjects who are older than thirteen years, it is necessary to refer to the full correction table provided in the manual. Or the examiner may use the tables of IQ's provided in the manual, in which the necessary adjustments have already been made.

^a From Terman and Merrill, *op. cit.*, p. 36. (By permission.)

DISTRIBUTION OF IQ'S

The mean IQ's, for the subjects used in the standardization, are slightly above 100. But this, the authors say, is due to an "intentional adjustment to allow for the somewhat inadequate sampling of subjects in the lower occupational classes." The adjustment was made by dividing the subjects into seven groups, according to the occupation of the fathers; at each age level the mean IQ was computed sepa-

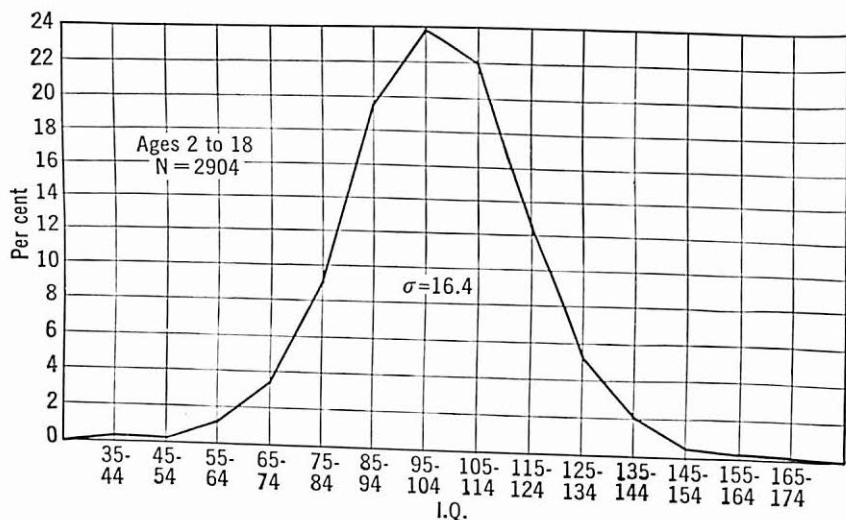


FIG. 6.2. Distribution of Composite L-M IQ's of Standardization Group. Terman and Merrill, *Measuring Intelligence*, Boston: Houghton Mifflin, p. 37. (By permission.)

rately for each of the seven groups. These means at each age level were given a weight according to the occupational frequencies of each group, as shown by the 1930 census. The weighted means were then combined into a composite mean for each age level from 2 to 18 years, as shown in Table 22. The same data are represented graphically in Figure 6.2.

In determining the equality and comparability of IQ's from age to age, it is necessary not only that the means be very much the same (ideally, identical), but that the *variations* be the same at all age levels. If the differences between the variations of the age groups are

large, then the same numerical IQ will have different significance at different chronological ages.

Consider the following hypothetical instance. Suppose a given test of mental ability yields the following results:

Chronological Age	Mean IQ	Standard Deviation
10	100	14
11	100	20

Accordingly, a 10-year-old child having an IQ of 86 (that is, one standard deviation below the mean) would have a percentile rating of approximately 16—which, it will be recalled, means that this child surpasses about 16 percent of his age group. Now, according to the foregoing data, a child of 11 years whose IQ is 80 (likewise one standard deviation below the mean of *his* group) would also have a percentile rating of about 16, in spite of the fact that his intelligence quotient is six points below that of the 10-year-old in question. While this difference of six points may make little practical difference in the clinical and educational treatment of these children, it is necessary to be familiar with the implications of differences in variations.

Another aspect of the problem is this. Taking our hypothetical case of the 10-year-old boy, if he is to have at age 11 the same *relative rank* he held at age 10, his IQ will have to drop from 86 to 80; and if this happens, the change will give the appearance of lack of IQ constancy; but, if he maintains his 86 IQ at age 11, then his percentile rank will be approximately 24. Here, again, the difference in percentile ranks might have little practical significance; but insight into the effects of differences in group variations will help explain some instances of lack of close agreement in repeated mental measurements. *Furthermore, the professional worker who uses mental tests must know the standard deviations of the instruments he employs in order to make as accurate an evaluation of his results as possible.*

IQ variability in relation to age, as found in the standardization of the 1937 Stanford-Binet is shown in Table 23. Inspection of this table shows significant fluctuation in standard deviations of the several age groups, especially between the extremes: namely, 12.5 (S.D.) at age 6, 20.6 (S.D.) at age 2½, and 20.0 (S.D.) at age 12. It will be noted, too, that the standard deviations fluctuate around 16 and 17 as a median value. The standard deviation of the composite IQ's (Forms L and M) is 16.4 for the entire standardization group of subjects.

In respect to the fluctuations in IQ variability, the authors state: "Notwithstanding our strenuous efforts to correct for . . . errors of sampling, complete success is hardly to be expected, and a considerable degree of irregular fluctuation in the found magnitudes of IQ variability from age to age could reasonably be attributed to these sources of error. . . . Since inspection of the values reveals no

TABLE 23
IQ Variability in Relation to Age ¹⁰
Stanford-Binet

CA	N	S.D. Form L	S.D. Form M
2	102	16.7	15.5
2½	102	20.6	20.7
3	99	19.0	18.7
3½	103	17.3	16.3
4	105	16.9	15.6
4½	101	16.2	15.3
5	109	14.2	14.1
5½	110	14.3	14.0
6	203	12.5	13.2
7	202	16.2	15.6
8	203	15.8	15.5
9	204	16.4	16.7
10	201	16.5	15.9
11	204	18.0	17.3
12	202	20.0	19.5
13	204	17.9	17.8
14	202	16.1	16.7
15	107	19.0	19.3
16	102	16.5	17.4
17	109	14.5	14.3
18	101	17.2	16.6

marked relationship between IQ variability and CA over the age range as a whole, we may accept 16 points as approximately the representative value of the standard deviation of IQ's for an unselected population."¹¹ As evidence to justify this position, the authors of the scale

¹⁰ From Terman and Merrill, *op. cit.*, p. 40. (By permission.)

¹¹ *Ibid.*, pp. 39-40. See also M. Aborn and G. F. Derner, "IQ Variability in Relation to Age on the Revised Stanford-Binet," *Journal of Consulting Psychology*, Vol. 15, 1951, pp. 231-235.

present the graph shown in Figure 6.3. These distribution curves of composite (L and M) intelligence quotients indicate that their variability is approximately the same for the three age-level groupings.¹²

Proceeding on the basis of the foregoing reasoning, that the standard deviation of IQ's is 16 points and that IQ values are comparable at all age levels, Terman and Merrill provide a table giving intelli-

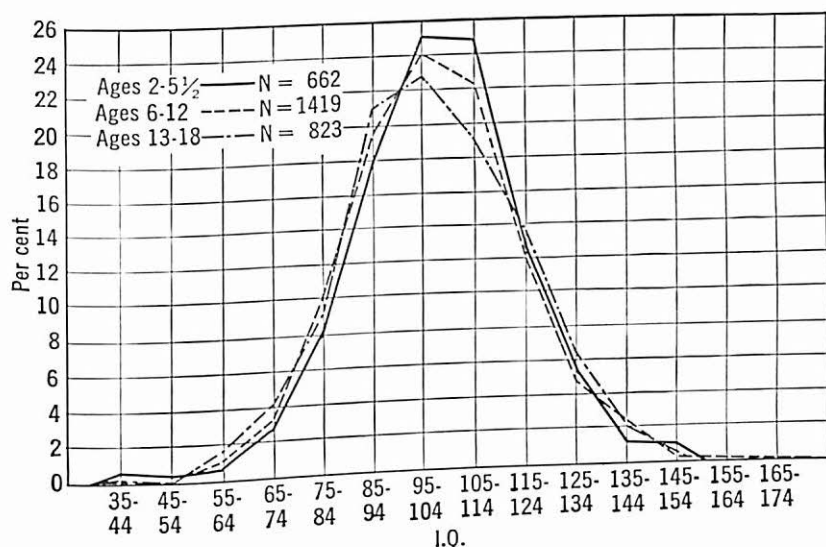


FIG. 6.3. *Distribution of Composite L-M IQ's at Three Age Levels.* Terman and Merrill, *Measuring Intelligence*, Boston: Houghton Mifflin, p. 41. (By permission.)

gence-quotient equivalents of standard scores (Table 24). This table, if Terman and Merrill's assumption of a "normal distribution" is accepted, is useful in giving more precise expression to the *relative* significance of any given intelligence quotient; for by means of an ordinary table of standard deviation values, it is possible to determine the approximate percentile status of any individual in the distribution of intelligence quotients.

¹² Some critics have not accepted the argument and conclusions of Terman and Merrill. It is our purpose, at this stage, only to present the scale and the rationale upon which it was constructed by its authors.

TABLE 24

*IQ Equivalents of Standard Scores*¹³

Stanford-Binet			
Standard Score	IQ	Standard Score	IQ
+5.00	180	— .25	96
+4.75	176	— .50	92
+4.50	172	— .75	88
+4.25	168	—1.00	84
+4.00	164	—1.25	80
+3.75	160	—1.50	76
+3.50	156	—1.75	72
+3.25	152	—2.00	68
+3.00	148	—2.25	64
+2.75	144	—2.50	60
+2.50	140	—2.75	56
+2.25	136	—3.00	52
+2.00	132	—3.25	48
+1.75	128	—3.50	44
+1.50	124	—3.75	40
+1.25	120	—4.00	36
+1.00	116	—4.25	32
+ .75	112	—4.50	28
+ .50	108	—4.75	24
+ .25	104	—5.00	20
.00	100		

TABLE 25

Distribution and Classification of Composite L-M IQ's of the Standardization Group

IQ	N	Percent	Classification
160-169	1	0.03	} Very superior
150-159	6	0.2	
140-149	32	1.1	
130-139	89	3.1	} Superior
120-129	239	8.2	
110-119	524	18.1	} High average
100-109	685	23.5	
90-99	667	23.0	} Normal or average
80-89	422	14.5	
70-79	164	5.6	} Borderline defective
60-69	57	2.0	
50-59	12	0.4	} Mentally defective
40-49	6	0.2	
30-39	1	0.03	

¹³ From Terman and Merrill, *op. cit.*, p. 42. (By permission.)

SUGGESTED CLASSIFICATION OF REVISED STANFORD-BINET IQ'S

The classification in Table 25 has been provided by one of the authors of the 1937 revision.¹⁴ It will be noted that the nomenclature and the percents in each of the several categories differ some from those of the 1916 instrument.

Like all such tables, its purpose is primarily descriptive and, also, to serve as an aid in the ordering and analysis of testing results. The table is valuable, as well, in showing an approximate distribution of intelligence quotients throughout most of the range of mental ability.

ANALYSIS OF FUNCTIONS TESTED

The items of both forms L and M have been analyzed by factorial methods. McNemar, who made the first and major analysis, concluded that at each of the several age levels the items test ("are saturated with") a common factor (g), and that this common factor is the same one at all age levels (hence it may be called g). The *weight* of the common factor differs somewhat among the various age levels; but the common factor accounts, on the average, for about 40 percent of the differences (variance) in scores—hence for about 40 percent of the differences in performance among a group of testees.

The statistical results also suggest the presence of group factors at the following ages: 2, 2½, 6, 18, and possibly 7 and 11. These are second factors (group factors) that account for from 5 to 11 percent of the differences; while a third factor (another group factor) contributes from 4 to 7 percent. The group factors do not appear to be identical at all age levels; nor are they at all well defined as regards the psychological processes involved in them. Tentatively, however, McNemar suggests that several of these group factors, at different levels, might be called "memory for designs," "motor," "verbal." *The most definite and significant conclusion, however, is that one factor (g) is sufficient to account for the intercorrelations of test items, with the few exceptions noted.*

When complex and comprehensive ("molar" or "global") types of test items are used in a scale, as in the Stanford-Binet (and quite appropriately so), it is not surprising that attempts to isolate group factors through statistical analysis yield results that are indefinite and uncertain, and at best tentative. The reason is that these types of items,

¹⁴ From M. A. Merrill, "The Significance of IQ's on the Revised Stanford-Binet Scales," *Journal of Educational Psychology*, Vol. 29, 1938, pp. 641-651.

being complex, involve a number of psychological processes, organized in varying degrees and interacting. Group and specific factors will be found most clearly when the test items employed are fractionations and small segments of a whole pattern of mental functioning. But such fractionation can destroy "the whole" and can fail to reveal the kinds of mental operations with which the examining psychologist is often most concerned.

The identification of a general factor in a revision of the Binet scale should not occasion any surprise, for it will be recalled that Binet himself set out to develop an instrument that should test an individual's general intelligence by means of sampling a variety of mental activities which are manifestations of such intelligence. It appears, therefore, that contemporary statistical analyses, applied to the age scale, are confirming Binet's psychological insights.

It was found, also—as Spearman had shown in his earlier analyses—that the various kinds of test items differed as regards the extent to which they tested (are loaded with) the general factor. The following listing shows which items were found to have high loadings of the general factor, and which had low loadings.¹⁵

AGES 2 TO 4½

<i>High Loadings</i>	<i>Low Loadings</i>
Picture vocabulary	Block building: tower
Identifying objects by name	Block building: bridge
Response to pictures	Three-hole form board: rotated
Comparison: balls and sticks	Motor coordination
Comprehension	Copying a circle
Opposite analogies	Drawing a cross
Pictorial identification	Three commissions
Naming materials [used in making various objects]	Stringing beads

AGES 5 TO 11

<i>High Loadings</i>	<i>Low Loadings</i>
Pictorial likenesses and differences	Paper folding: triangle
Similarities: two things	Patience: fitting rectangles
Vocabulary	Copying a bead chain
Verbal absurdities	Copying a bead chain from memory
Similarities and differences	Picture absurdities
Naming the days of the week	Word naming [free association]
Dissected sentences	Word naming: animals
Abstract words [definitions]	Block counting

¹⁵ From McNemar, *op. cit.*, pp. 111-113.

AGES 12 TO SUPERIOR ADULT III

High Loadings

Vocabulary
 Verbal absurdities
 Abstract words [definitions]
 Differences between abstract words
 Arithmetical reasoning
 Proverbs
 Essential differences
 Sentence building

Low Loadings

Problems of fact
 Copying a bead chain from memory
 Memory for stories
 Enclosed box problem
 Paper cutting [visual imagery]
 Plan of search
 Repeating digits [forward]
 Repeating digits: reversed

Examination of the items that have low loadings reveals that they test only a very limited range of functioning and that, with two exceptions, they involve only the following processes: visualization (space perception and spatial relationships), visual imagery, and rote memory (immediate recall). All of the test items, among the low loadings, falling under the foregoing categories are lacking, relatively, in complexity and would, therefore, not have the differentiating power of the more complex tasks required by the items under high loadings. The exceptions are: word naming (random and animals), and problems of fact. The latter is properly considered to be a test of reasoning. Its low loading with *g* is therefore surprising, but no explanation is apparent or available. Random word naming tests richness of free association; word naming of animals tests controlled association. A possible explanation of their low loadings might be that they are fairly routine tasks that do not require the reasoning (organization, analysis) demanded by the items having high loadings within the same range of ages (5-11).

The foregoing listings are significant for at least two additional and important reasons. First, a knowledge of test items that have high or low loadings of the general factor enables the examining psychologist to make a more thorough analytical and meaningful evaluation of an individual's over-all test performance. The examiner is thus in a better position to evaluate the strength of the general factor in a particular testee. This is particularly valuable if the psychological nature of the general factor has been determined. Second, inspection of the lists of items having high loading strongly indicates that the general, or common, factor is one that involves acquisition of, use of and reasoning with symbols—namely, language and number—even though the testing of these begins at a very elementary level and at times utilizes nonverbal materials in presenting the problem (e.g., pictures, sticks).

The mental activities required by these test items have very much in common with Spearman's view that intelligence is essentially the ability to educe relations and correlates.

Specifically, the following processes are involved in the test items having high loadings: acquisition and use of vocabulary; verbal analysis of a situation; verbal and numerical concept formation; insights into similarities and differences (also involving concept formation); analysis and synthesis of materials, both nonverbal and verbal; organization and reorganization of materials, both nonverbal and verbal.

The list of items given above (based on statistical calculations) and the indicated psychological functions that are involved provide an illustration of how statistical and psychological analyses work together. They also make it clear that superficial observations of differences between test items can be misleading as to their essential psychological processes. For example, the test items at early age levels requiring identification of objects by name or use might be regarded simply as tests of information or of specific rote learning, whereas they actually have much in common with items that test "comprehension," and which are more obviously tests of reasoning. Or, at a somewhat later age level, ability to define certain words (vocabulary test) might be regarded simply as the result of specific learning and verbal facility, whereas actually it has much in common with perception of pictorial (as well as verbal) similarities and differences.

It is useful to classify test items as "information," "word knowledge," "perception of forms," "reasoning," etc.; but the point is that such classification does not necessarily signify that each of the subtest classifications measures a distinct group-factor or a special factor.¹⁶

¹⁶ McNemar's findings are in close agreement with those of an independent study made in Great Britain. Cyril Burt reports that a common factor accounts for 42 percent of Stanford-Binet test variance, and that two subsidiary factors account, respectively, for 12 percent and 16 percent of test-score differences. The close correspondence obtained in the United States and in England gives additional weight to and confidence in the Stanford-Binet Scale as an instrument for measuring intelligence, particularly of children and adolescents. See Cyril Burt and Enid John, "A Factorial Analysis of Terman Binet Tests; Part I, *British Journal of Educational Psychology*, Vol. 12, 1942, pp. 117-127; *ibid.*, Part II, Vol. 12, 1942, pp. 156-161.

For a dissenting analysis and interpretation, see L. V. Jones, "A Factor-Analysis of the Stanford-Binet at Four Age Levels," *Psychometrika*, Vol. 14, 1949, pp. 299-331.

TYPES OF ITEMS

Bearing this important distinction in mind, then, we may indicate the types of items included in the Stanford-Binet Scale.¹⁷

<i>Test Items</i>	<i>Functions Involved</i>
<i>Years 2-5</i>	
Form perception and manipulation (blocks, form boards, stringing wooden beads)	Visual perception and analysis
Perception of differences in size and form	
Visual-motor operations	Visual analysis plus motor development
Perception of relationships (in pictures)	Visual perception plus beginnings of concept formation
Rote memory (using digits and sentences)	Immediate recall
Use of words in combination	
Identifying objects by name or use	Language development and comprehension
Following directions	
Verbal comprehension and word knowledge	Reasoning with abstractions and concept formation
Understanding of "opposites"	
<i>Years 6-12</i>	
Form perception	Visual analysis
Visual-motor operations	Visual analysis plus motor development
Rote memory (using digits and sentences)	Immediate recall
Word knowledge (concrete and abstract)	Language development

¹⁷ Since the Bellevue Scale utilizes many of the same types of items as the Stanford-Binet, additional factors affecting test performance on these items will be presented in connection with the Bellevue Scale in the next chapter.

<i>Test Items</i>		<i>Functions Involved</i>
Verbal comprehension		Reasoning with abstractions and concept formation
Number concepts		Number concept formation
Arithmetical reasoning		tion and reasoning with abstractions

Year 13—Superior Adult III

Visual analysis and imagery		Visual perception and
Perception of visual relations		analysis plus reasoning with non-verbal materials
Visual-motor operations		
Rote memory (using digits, words, and sentences)		Immediate recall
Word knowledge		Language development
Synthesis of verbal materials		
Problem solving, using verbal materials		Reasoning with abstractions
Verbal analysis		
Arithmetical problems		Concept formation plus
Analysis and comprehension of symbols		reasoning with abstractions

THE SHORT SCALE

It is possible to administer an abbreviated form of the scale, the constituent items having been specified by the authors. A short scale, presumably, is used when the examiner does not need as accurate an index of measurement as it is possible to obtain and when the necessary time is not available for a full-length examination. The use of an abbreviated form, however, should be discouraged, for when it is at all desirable to administer a test of mental ability, it would be very unwise indeed not to require the greatest possible accuracy.¹⁸

¹⁸ For some results obtained with the short scale, see, for example, G. Spache, "Methods of Predicting Results of Full-Scale Stanford-Binet," *American Journal of Orthopsychiatry*, Vol. 14, 1944, pp. 480-482; G. Spache, "The Abbreviated Stanford-Binet Scale in a Superior Population," *Journal of Educational Psychology*, Vol. 35, 1944, pp. 314-318.

EVALUATIONS AND CRITICISMS

Every scale, quite properly, is subjected to evaluation and criticism on both theoretical, practical, and experimental grounds. The 1937 Stanford-Binet is no exception. On the whole, educators and clinicians who have had experience with both the old and the new Stanford-Binet scales are in essential agreement that the new one, in spite of some inadequacies, is a more useful instrument of its kind than the older one. The following, however, are some of the questions and criticisms that have been raised.

Is the age-scale (Binet) type of test preferable to the point-scale type of test? The standardization of an age scale is much more laborious and rigid than standardization is in the case of a point scale. The result is that even when experience and experimentation reveal certain defects and inadequacies in an age scale, the difficulties in making the indicated changes are so great as to be a deterrent to early revision.

As instances in point, some psychologists report difficulties in interpreting responses to certain items and in scoring them. They also question the age placement of some items. The task of correcting these defects, if the criticisms are warranted, would be great.¹⁹ Note, for example, that the authors of the 1937 Stanford-Binet made six revisions of Form L before they were sufficiently satisfied with the age placement and grouping of test items, which would yield correct mental ages, intelligence quotients, and IQ distributions.

In the case of a point scale, on the other hand, it is a much simpler process to revise age norms. All other things being equal, of course, the simpler and easier methods should be employed to achieve a desired goal in psychological testing. But simplicity and ease alone should not be the decisive considerations. The crucial question is whether the age-scale type of test or the point-scale type provides a superior means of obtaining a measure of an individual's mental ability. Thus far, although the views of competent professional persons are not unanimous, it appears that the age scale is preferable for use with children and young adolescents. But when older adolescents

¹⁹ For example, M. Krugman, "Some Impressions of the Revised Stanford-Binet Scale," *Journal of Educational Psychology*, Vol. 30, 1939, pp. 594-603; H. E. Garrett, "The Standardization of the Terman-Merrill Revision of the Stanford-Binet Scale," *Psychological Bulletin*, Vol. 40, 1943, pp. 194-201.

and adults are to be examined, it has often been found that point scales (e.g., the Wechsler-Bellevue) are preferred.²⁰ Also, there is at present a trend toward increased use of point scales because of their apparent value in clinical diagnosis by means of "scatter analysis" of scores on the subtests. (See Chapter 15.)

Do the differences in variability (standard deviation) at the different age levels seriously impair the usefulness of the Revised Stanford-Binet Scale? A wholly satisfactory test of intelligence should have equal or very nearly equal variability at all age levels.²¹ Data already

TABLE 26
IQ's Adjusted for Variability Differences
at Several Age Levels²²

Obtained IQ's	Adjusted IQ's			
	2-4 to 3-3	4-10 to 6-6	11-6 to 12-5	14-6 to 15-5
140	134	149	135	136
130	125	137	126	127
120	117	125	117	118
110	108	112	109	109
100	100	100	100	100
90	92	88	91	91
80	83	75	83	82
70	75	63	74	73
60	66	51	65	64
50	58	39	56	56

cited (Table 23) show that the standard deviations of IQ's as found by Terman and Merrill were not equal at all age levels, though most fell close to 15-17 points. However, data have been provided to adjust for variability differences in IQ at age levels where such adjustments are necessary. Table 26 provides some examples of these adjustments.

Table 26 is read thus: If an individual's age is between the limits of 2-4 and 3-3 and his obtained IQ is 140, his adjusted rating would be 134.

Two aspects of this table should be noted in particular. First, the

²⁰ For example: F. Halpern, "A Comparison of the Revised Stanford L and the Bellevue Adult Intelligence Test as Clinical Instruments," *Psychiatric Quarterly Supplement*, Vol. 16, 1942, pp. 206-211.

²¹ This requisite is based upon the principle that the distribution of general ability is equal at all the age levels covered by the test, rather than being either irregular from year to year or changing systematically at successive ages.

²² From Q. McNemar, *op. cit.*, pp. 173-174. (By permission.)

adjustments are small for IQ's near the average (100) and larger as the obtained IQ's deviate more from the average. Second, in no instance would the adjusted IQ seriously displace an individual from one *general* level of ability to another, even though in a few instances the changes are appreciable, notably from an IQ of 50 to one of 39 in the 4-10 to 6-6 range.

The answer to our question is, then, that differences in obtained IQ variability on this scale do not seriously impair the usefulness of the Revised Stanford-Binet Scale; and they *need not* impair its usefulness if the user of the scale is aware of the differences in variability at the several age levels and makes the necessary adjustments of IQ values. It is to be expected, of course, that later revisions of the Stanford-Binet will minimize the adjustments.²³

Are the Stanford-Binet test items a variety of disconnected tests? This scale has been criticized at times as being only that. This criticism, however, is based upon a failure to take into account the theory of intelligence and the method of measurement upon which the scale is based: namely, the sampling of *general ability* by means of a wide variety of types of items in order to obtain an adequate estimate of the processes involved. Furthermore, the factorial analyses already discussed in this chapter show that the test items measure primarily a general factor that is common to all age levels of the scale.²⁴ On cursory inspection, the items may *seem* to be dissociated; but psychological and statistical analyses demonstrate that such is not the case in fact.

Is the composite, or "global," type of scale (such as the Stanford-Binet) preferable to the factorial analyzed type, in which "factors" or "primary abilities" are separately tested and scored? Again, the final answer to this question will depend upon whether the factorial type

²³ The cause of the differences in standard deviations in IQ at different age levels may be due to either or both of the following: (1) unequal item difficulty at the several age levels; (2) inadequate samplings of the standardization population. See McNemar, *op. cit.*, Chapter 8; also A. L. Baldwin, "Variation in Stanford-Binet IQ Resulting from an Artifact of the Test," *Journal of Personality*, Vol. 17, 1948, pp. 186-198; J. A. F. Roberts and M. A. Mellone, "On the Adjustment of Terman-Merrill IQ's to Secure Comparability at Different Ages," *British Journal of Psychology, Statistical Section* 5, 1952, pp. 65-79.

²⁴ Even the analyses that emphasize group factors rather than a general factor do not support the criticism that the items are merely "a variety of disconnected tests."

of scale proves to be more valid, more accurate, and more useful than the composite type has been in clinical work and in educational and vocational guidance.

It is not improbable, however, that psychologists will develop satisfactory tests of intelligence that will yield an index of general ability (e.g., mental age and intelligence quotient) and which, at the same time, can be analyzed and scored for particular aspects of intelligence or particular abilities. To do this, instead of grouping items by age level, it will be necessary to devise a number of parts, or subtests (e.g., verbal reasoning, numerical ability, etc.), each of which would contain a series of items, all items within each subtest measuring the same mental processes, but scaled in difficulty. Each subtest would yield a separate index of test performance; and at the same time the individual's average performance would represent his general level. This is what the Wechsler scales (discussed in the following chapter) aim at.

At the present time, the Binet type of test provides only the general indexes of MA and IQ. But—and this matter will be dealt with more fully later—the qualified examiner is able to make a significant *qualitative* analysis of a subject's performance on the scale.

Is the Stanford-Binet Scale too heavily weighted with verbal materials?

A criticism, heard with some frequency against both the old and the new revisions, is that these scales place a premium upon "verbal intelligence" and that subjects having language handicaps are penalized and incorrectly rated. In reply to this criticism, Terman and others hold that the most essential and most significant aspect of higher thought processes is the ability to do conceptual and abstract thinking; that is, to operate with language, number, and other symbols. It is maintained, also, that the vocabulary test, when used with children from homes where English is the primary language, has higher value than any other part of the scale.

It must be emphasized, however, that this is not so in the case of a child who, even though he comes from such a home, has reading or language difficulties due not to lack of capacity but to visual or auditory defects or anomalies.

In actual clinical practice, the examiner should always supplement an essentially verbal test of intelligence with one of the nonverbal type if he has any reason to suspect that the former penalizes the subject. There will be occasions also when it will be desirable to obtain a

rating for an individual on both types, even where no language handicap is indicated, for the purpose of comparing two or more aspects of a subject's ability. While the correlation between performance on verbal and on some nonverbal tests of mental ability is high, coefficients of correlation reflect *group* trends and relationships; and unless the correlation coefficient is perfect (plus or minus 1.00), there are always individual exceptions from the generalization that can be made on the basis of the coefficient; hence the need, at times, for the study of the several aspects of ability in an individual case.

Is the Stanford-Binet Scale a test of school learning? As stated in an earlier chapter, any scale must test mental ability through its manifestations, through activity of some kind. And as Binet originally pointed out, the tests must be adapted to the environment of the subjects to be rated. It is reasonable and sound, therefore, that a scale designed to test mental ability primarily of American children and adolescents should utilize the *effects* of common schooling experiences, as well as effects of some other common experiences. To say that the Stanford-Binet or any other scale is only a measure of school learning is unwarranted. Differences in quality and extent of opportunity to learn, in school and out, will have an effect upon intelligence-test scores; but such differences in opportunity do not in themselves account for all individual differences in ability that are found. Good schooling and other good environmental conditions nurture an individual's mental capacities and provide optimal conditions for his mental development. Thus we can say that the Stanford-Binet and other scales provide ratings of intelligence, within limits of error, under existing school conditions, general environmental conditions, and clinical conditions. Obviously, therefore, the examiner must know the general developmental background of the individual he is testing, if his interpretation of test results is to have real validity.

Several studies of satisfactory and unsatisfactory responses on the Stanford-Binet have found that bright and superior children answer correctly more of the "intellectual" items than do normal or dull children.²⁵ These items include the verbal and numerical, utilizing symbols

²⁵ See A. L. Baldwin, "The Relative Difficulty of Stanford-Binet Items and Their Relation to IQ," *Journal of Personality*, Vol. 16, 1948, pp. 417-430; A. Magaret and C. W. Thompson, "Differential Test Responses of Normal, Superior, and Mentally Defective Subjects," *Journal of Abnormal and Social Psychology*, Vol. 45, 1950, pp. 163-167.

and abstractions. This finding does not mean that bright and superior children have attained their ratings on mental tests only as a result of schooling. It is to be expected that individuals who are potentially above average in mental ability will be superior in dealing with situations and problems employing language and number; for the greater the capacity of the individual is for mental development, the greater will be his ability to deal with symbols and to handle situations and problems at the level of abstraction. The converse has also long been known, namely, that one of the principal deficiencies of the mentally retarded and mentally defective is their inability to deal with materials and concepts at the levels of abstraction. It will be recalled that one definition states that intelligence is the ability to deal with abstractions. It will be recalled, also, that the educing of relations and correlates extends upward to the use of symbols (language and number).

Are some of the items in the Stanford-Binet Scale obsolete? Since any test utilizes materials from the environment in which it is to be used, and since environments normally undergo change, it is to be expected that some items in any given test will in time become culturally obsolete. In the Revised Stanford-Binet there are a few such items. For example:

- Identifying a toy steam locomotive by name
- Identifying objects (pictures) by use, such as an old-fashioned kitchen stove
- Response to a picture (Messenger Boy, year 12)

There are not many such items in this scale. In the case of some of the non-obsolete items, however, it becomes necessary, with time, to revise to some extent the responses that are acceptable for credit. For example, the following item is one such: "What's the thing for you to do when you are on your way to school and see that you are in danger of being late?" (Year 7) When scoring an unusual response to an item like this, the qualified examiner is warranted in exercising his judgment as to its correctness; in fact, he has to do that. And his decision will be based upon his familiarity with the psychological processes being tested by the particular item.

Does the Stanford-Binet Scale test different abilities at different age levels? The answer to this question is to be found largely in the preceding discussion of "Analysis of Functions Tested" in this chapter.

There it was stated that the two major analyses (by McNemar and Burt) found that the scale measures principally a general factor that is common to all age levels, and in addition there appear to be group factors at 2, 2½, 6, 18, and possibly at 7 and 11. It was pointed out that the items having low loadings of the general factor were, with minor exceptions, tests that required the use of visual perception, visual imagery, and rote memory. Emphasis was placed, also, upon the necessity of distinguishing between item-type classifications of tests (vocabulary, arithmetical reasoning, etc.) and basic psychological processes involved in each.

Does the scale measure originality and creative abilities? The answer is that it does not measure these abilities, as such, to an important degree. This aspect of intelligence was discussed in Chapter 3, "Definitions and Analyses of Intelligence." There it was pointed out that the requirement of *objectivity* in standardizing and scoring tests practically excludes tests of creativeness and originality. But, as also stated, while not all individuals who achieve high scores on intelligence tests evidence originality and creative ability, those who are capable of originality and creative mental activity do generally obtain high test scores. It is possible to say, therefore, that persons having creativeness and originality will be found generally in the group who attain superior ratings on the intelligence scales. Furthermore, although originality and creativeness cannot be included in the prescribed objective scoring, a qualified examiner will note responses indicating these traits and will include them in his interpretation and evaluation of the examinee's performance.

Is the 1937 Stanford-Binet Scale adequate at the adult level? The standardization group of this scale included individuals of 18 years, but it did not include an adult population. Therefore, the test items at the several adult levels rest upon theoretical considerations already mentioned, rather than upon actual samplings of adult performance. One result, due perhaps to the methods used in standardizing the scale at superior adult levels, has been its frequently observed inadequacy with college students who, as a group, would be ranked above average. The inadequacy of the scale is especially marked when administered to very superior students, for it is not difficult enough at the higher levels of adulthood.

Some psychologists have also questioned whether some of the items and materials employed at the adult levels are of sufficient interest to adults. These critics believe there are a number of items that are unsuited to the maturity level of adulthood. Finally, there is the question of the soundness of using the mental-age index for adults above the mental level of average adult. The reason for this question has already been presented. Many psychologists, therefore, prefer to discard the mental-age index at adult levels, in favor of percentile ranks, standard scores, and the like.

Is the 1937 Stanford-Binet Scale clinically useful? Judging by the extent to which it is employed, the answer must be a strong affirmative. Examiners find the scale useful not only for deriving a mental age and an intelligence quotient, but also as the "framework" within which a psychological interview is held. If the scale is to serve this purpose, the examiner must have considerable clinical experience and skill to interpret and evaluate a subject's responses and behavior.

In spite of its wide clinical use, however, there is some difference in evaluations of the instrument among clinical examiners. Kent takes the position, for example, that the scale is too complex and too rigid.²⁶ She would have a graded series of items for each of the several types of tests. The clinician would then use a "battery"—or selection—of these, "custom-made" for each individual, to meet each special clinical problem. Separate norms and relative ratings for each graded series would have to be provided, from which median performance would be determined.

This criticism and proposal are countered by Vernon, for instance, who states that there is great value to a clinician in having available objective and standardized measuring devices, such as the Stanford-Binet. This scale, it is maintained, provides the clinician with an excellent psychometric device which, at the same time, permits sufficient flexibility to meet the situation as may be necessary in a particular case.²⁷ Many experienced psychologists agree with Vernon's position.

²⁶ G. H. Kent, "Suggestions for the Next Revision of the Stanford-Binet Scale," *Psychological Record*, Vol. 1, 1937, pp. 409-433; see also V. V. Fleming, "A Study of the Subtests in the Revised Stanford-Binet Scales, Forms L and M," *Journal of Genetic Psychology*, Vol. 46, 1944, pp. 3-36.

²⁷ P. E. Vernon, "The Stanford-Binet Test as a Psychometric Method," *Character and Personality*, Vol. 6, 1937, pp. 99-113. See also G. W. Parkyn, "The Clinical Significance of IQ's on the Revised Stanford-Binet Scale," *Journal of Educational Psychology*, Vol. 36, 1945, pp. 114-118. This article is concerned mainly with individuals having IQ's under 80.

As new scales are devised—built upon different conceptions and theories—they will have to be subjected to both experimental investigation and practical use before their value can be compared with that of the Stanford-Binet. A valid judgment cannot be reached on *a priori* grounds. In the meantime, the great value of the Stanford-Binet Scale having been demonstrated, psychologists will continue to use it. They will bring to bear their insights on the interpretation of the *behavior* of the persons being examined and on the interpretation of the *test results* obtained.²⁸

²⁸ The reader will find a very useful analytical table giving the content of the several Binet scales and of the several American revisions in G. D. Stoddard, *The Meaning of Intelligence*, New York: Macmillan, 1943, pp. 483-497.

In a personal communication [March, 1954] Dr. Maude A. Merrill wrote: "Analysis of my preliminary sample of some 800 cases [tested within the last five years] between five and fourteen yields results that are very heartening. The indications are that the scale is doing in 1954 pretty much what it did in 1937, that there are some differences in difficulty in certain of the subtests and specifically certain of the items that can be changed in accordance with the indications by adding pretested substitute or modified items without doing violence to the original structure of the scale. The curves showing increase in percent passing with age are remarkably regular and follow very closely the original percentages except on a rather surprisingly small number of items. One finding . . . was that the modern elementary-school child is handicapped on tests that assume the acquisition of reading skills. In 1937 we could assume that, on the average, by the age of ten, children had acquired reading techniques that would enable them to make use of such skills in the problem-solving tasks presented by certain of the subtests. *Reading and Memory* at the ten-year level, *Minkus Completion* at the twelve-year level and *Dissected Sentences* at thirteen stand out . . . as presenting much more difficulty than in 1937. The obvious inference is that this finding reflects differences in educational practices. These will need to be verified on another sample." (Reproduced with permission of Dr. Merrill.)

THE WECHSLER SCALES

THE Binet scale and its several revisions are largely verbal in content, although some nonverbal items are included, especially at the early age levels. There are, however, other scales which are wholly, or in large part, of the performance or nonverbal type. This type of test is one in which the use of language is eliminated from test content and response, although directions are generally given orally. In a few instances the directions, too, are given without the use of language, by employing pantomime instead.

The test materials of a nonverbal scale consist of concrete objects such as form boards, cubes (to be arranged in specified ways), mazes, geometric figures, pictures (cut up, to be correctly assembled), and others that will be described in later sections. The individual's responses take the form of manipulations, visual perceptions, and interpretations which are implied by what he *does* rather than by anything he says.

Performance tests were first devised as a supplement to or substitute for the Stanford-Binet scale in order to examine deaf, illiterate, or non-English speaking subjects. Since their introduction, the use of nonverbal tests has been extended; for they are now utilized with children who have or are suspected of having reading difficulties, with those who have attended school irregularly and might thus have been handicapped in developing verbal ability, and with persons who might have been handicapped by markedly inferior environmental conditions. Nonverbal tests are used, also, by examiners who, for any other reason, believe that such a scale will yield a more complete picture

of the individual whose capacities are being analyzed and evaluated.

The Wechsler scales, to be described in the following pages, combine verbal and nonverbal materials within a single instrument in an effort to obtain the advantages, comparisons, and contrasts yielded by both types of materials.

DESCRIPTION OF THE WECHSLER-BELLEVUE INTELLIGENCE TEST¹

This scale, published in 1939, is intended to test the intelligence of persons from the age of 10 through the age of 60 years, although norms are available beginning at 7.5 years. This or a similar beginning level is necessary, of course, if adults and adolescents of lower-than-average mental levels are to be tested by means of the scale.

The Bellevue scale will be presented in some detail so that the reader may understand the rationale of its construction, its values, and its limitations.

Content of the Scale. The scale is in part verbal and in part performance, enabling the examiner to obtain three scores and three quotients: a full-score and its intelligence quotient, a verbal score and its IQ, a performance score and its IQ.

The scale has been constructed in this form on the principle that intelligence involves not only ability to deal with symbols, abstractions, and conceptual thinking, but that it also involves ability to deal with situations and problems in which concrete objects, rather than words and numbers, are utilized. The scale also is put to the pragmatic test of whether a given combination of items (in this instance verbal and nonverbal) serves the purpose of individual mental examination and analysis of capacities better than other combinations. The types of tests included in this scale are not unique. They were selected from available sources after a study had been made of a variety of standardized tests then in use. The objective was the construction of an

¹ Cf. D. Wechsler, *The Measurement of Adult Intelligence*, Baltimore: Williams and Wilkins, 1944, third edition. A second form of this scale was published in 1946. Form II is identical with the first form in respect to underlying principles and types of test materials. The specific content, of course, is different. See *The Wechsler-Bellevue Intelligence Scale: Form II*, New York: The Psychological Corporation, 1946.

effective scale for adolescents and adults, based upon already known and proven psychological materials and procedures.

The Bellevue scale consists of eleven parts, or subtests, six being verbal in content and five nonverbal.² The principal difference between the Stanford-Binet and the Bellevue scale, in respect to arrangement of items, is this: in the former, items of various types, testing a

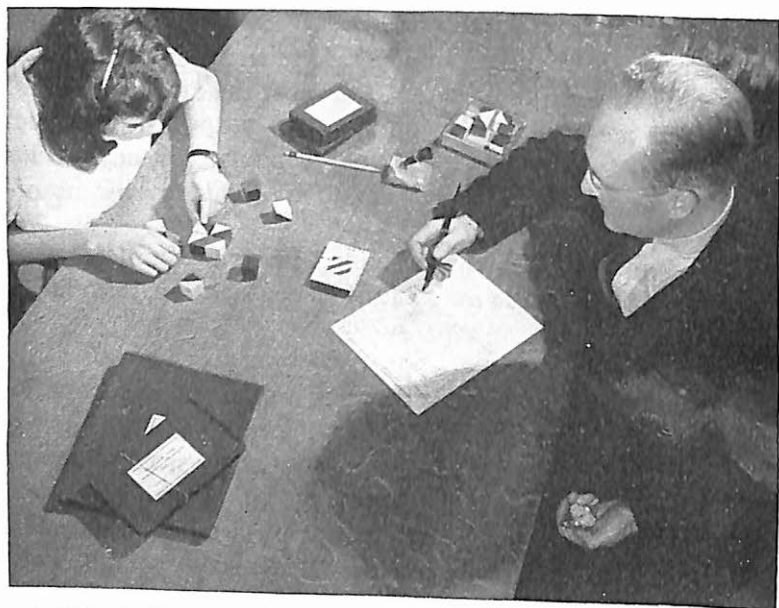


FIG. 7.1. A 12-year-old girl is shown here taking the block-design test of the Bellevue scale. The model of the design to be copied is on the card before her. The examiner is timing her with the stopwatch in his left hand. (Acme Photo.)

variety of functions, are grouped together at each age level; in the latter, all items of one type are grouped together, constituting a subtest of the whole. In the case of the latter, the effort is made to arrange the individual items within each subtest in a sequence of increasing difficulty. A description of the subtests follows. The first six constitute the "verbal scale"; the remaining five constitute the "performance scale."

² One of the verbal subtests, vocabulary, was originally added as an alternate or supplement. However, it is now used quite regularly with the other subtests.

(1) *Information.* This consists of twenty-five items of information, covering a wide range. (For example, "How many weeks are there in a year?") The assumptions are that the questions cover a wide enough range of materials to provide an adequate sampling of information acquired by a person who has had the usual opportunities of our society; that the range of an individual's information is an indication of his intellectual capacity; and that the more intelligent have broader interests, more curiosity, and seek more mental stimulation. *This view can be valid, however, only if the subjects being tested have had the usual opportunities for experience and learning and if the test items are a valid sampling of the opportunities to acquire information.* It is necessary to keep this caution in mind in evaluating a test of information. Performance on information tests is susceptible, also, to variations in individual motivation; i.e., some bright, self-absorbed persons or those unreceptive, for emotional reasons, to the offerings of their environments have unduly limited funds of information. Others, for motivational reasons quite removed from general intelligence, exhibit an exaggerated and misleading fund of information. For these very reasons, however, a test of information is a useful one for clinical purposes: that is, for purposes of diagnosing personality traits.

(2) *General comprehension.* This part of the scale consists of ten problem situations (plus two alternates), in which the subject must comprehend what is involved in the situations and provide answers to problems presented. (For example: "Why should people pay taxes?") Success on this subtest, it is held, depends upon possession of practical information plus ability to evaluate and utilize past experience. It appears, also, that ability to verbalize is a factor contributing to success. Tests of general comprehension are now very commonly used in intelligence scales; and, it will be recalled, they were included in the Binet scales. They have been found valuable clinically in revealing the thought processes, background, feelings, and emotions of the subject.

(3) *Arithmetical reasoning.* This part of the Bellevue scale is designed to test "mental alertness." The problems, the author of the test states, do not require "knowledge" (presumably arithmetical skills) beyond that of the seventh-grade level. Problems in arithmetical reasoning are widely used in tests of intelligence, since they have been found to have significant correlation with total scores of scales and to have high predictive value in respect to future evidences of mental ability.

(4) *Memory span for digits, forward and backward.* The subject is required to repeat series of digits heard once. The series vary in length from three to eight (backward) and nine (forward). This is a test of immediate recall, or immediate memory span. Psychological studies—both experimental and clinical—have consistently shown that tests of memory span, or immediate recall, of digits have a low correlation with other, more valid tests of intelligence. Yet, memory span for digits continues to be used because it is helpful in detecting the mentally defective, whose span is often very short (generally less than five digits forward and less than three backwards), and because very poor span is useful in making certain clinical diagnoses of organic defects. Poor memory span for digits, especially backwards, is also found at times in cases of persons who are unable to apply the attention necessary in solving more difficult mental tasks.

(5) *Similarities.* This part of the scale consists of twelve sets of paired words; for each pair the subject is required to state the similarity or similarities that exist. (For example, orange-banana.) The author of the scale regards the similarities test as one of the most satisfactory, for it appears to sample very well the "general factor" (Spearman's *g*), or what is commonly called *general intelligence*.

(6) *Vocabulary.* This test consists of forty-two words selected from an original list of one hundred which were chosen from a dictionary according to a sampling formula³ and then experimentally arranged in order of difficulty. The reader is already familiar with the view held by many psychologists that a vocabulary test—where there have been no unusual developmental factors—is one of the most valuable kinds of materials in deriving an index of a person's general intellectual ability. Thus, although the vocabulary list was originally recommended as an alternate test in the Bellevue scale, experience demonstrated its value, so that it is now suggested that this part be included regularly when the full scale is to be administered. Also, like Binet and many other psychologists, users of the Bellevue scale observed that qualitative differences in word definitions, given by various subjects, have clinical value in helping to reveal the nature of an in-

³ "The words . . . were taken from one of the Funk and Wagnall's Standard (School) Dictionaries. The list was arrived at by choosing 100 words at random in the following manner: Beginning with an odd page, we selected every top word but one (omitting, however, obsolete, technical, or esoteric words) in the left-hand column of every fifth page and continued the process until we had 100 words." (Wechsler, *op. cit.*, p. 99.)

dividual's thought processes (their depth, extent of analysis, nuances of meanings, queerness of definitions, cultural background) and, in some instances, feelings, emotions, and values.

(Numbers 6 through 10 are the performance tests.)

(7) *Picture arrangement.* In this subtest of the Bellevue scale, there are seven series of pictures. Each series is presented to the subject in a disarranged order; but when the pictures in each series are placed in the correct sequence, they tell a story. This type of test, it is held, measures a person's ability to comprehend and evaluate a total situation without the use of language. From the author's data, however, it appears to involve the "general factor" only to a moderate degree, but sufficiently to make a contribution to the total sampling.

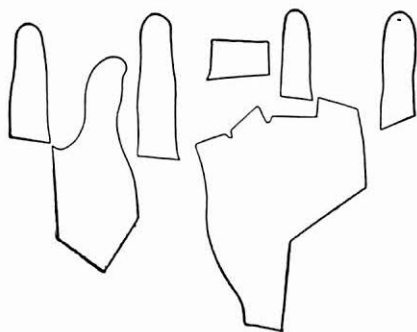


FIG. 7.2. *The Disassembled Hand (Object-Assembly Test).* From the Wechsler-Bellevue scale. (By permission.)

(8) *Picture completion.* In this part, there are fifteen cards, each of which shows a picture that is incomplete in some detail (for example, a picture of a face with the nose missing). The testee is required to note and name the missing part.⁴ In some pictures the task is quite simple for the ordinary person; but in others the deficiencies of the pictures are somewhat more subtle. It has been found that this material is particularly valuable in testing lower-level intelligence, as well

⁴ Generally this type of test is called "mutilated pictures." It was used by Binet in his scales and is now widely used in group tests, as well as in Binet revisions. The term *Picture Completion* is misleading, since the testee does not actually complete the picture.

as having moderate discriminative value at the intermediate levels. At the higher levels, however, this test is inadequate because it is not of sufficient difficulty. On the whole, it is said that this part of the scale ". . . measures the individual's basic perceptual and conceptual abilities in so far as these are involved in the visual recognition and identification of familiar objects and forms. . . . In a broad way, the test measures the ability of the individual to differentiate essential from nonessential details."⁵

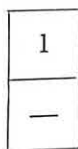
(9) *Block design*. This test utilizes sixteen identical cubes, some or all of which are used to make nine given designs (two of which are for demonstration). One side of each cube is colored blue, one red, one white, one yellow, one red and white, one blue and white (the last two being divided diagonally). For each of the four easiest designs, only four blocks are used; for the next two, nine blocks are used; for the most difficult design, sixteen are used. The author of the scale believes that this test involves ability to analyze and synthesize. He reports that scores on block design correlate well with total scores of the test, as they do also with the separate scores on the comprehension, information, and vocabulary tests of the Bellevue scale. This would indicate that the block-design test is valuable as a measure of the general factor.

(10) *Object assembly*. This test includes three "figure form-boards"; that is, three familiar objects made of wood, each one cut into several parts which have to be assembled to make the whole. The objects are a *manikin*, a *feature profile* (side view of a human head), and a *hand*. The inclusion of this test is justified by the author on the ground that it is desirable to have a device which requires the subject to perceive and reconstruct the parts of familiar objects into their wholes. Evidence on the object assembly tests indicates that it correlates poorly with other parts of the scale and has only very limited value in differentiating between individuals. It does, however, have clinical value, of a qualitative kind, in that it contributes to the examiner's understanding of the subject's "modes of perception, the degree to which one relies on trial and error methods, and the manner in which one reacts to mistakes."⁶ This subtest is useful, also, in diagnosing disturbance of visual perception.

⁵ Wechsler, *op. cit.*, pp. 90-91.

⁶ Wechsler, *op. cit.*, p. 98.

(11) *Digit-symbol test*. The subject is shown nine divided rectangles; in the upper half of each rectangle is a digit; in the lower half there is a symbol. For example, the following:



The key is followed by 75 rectangles (of which eight are practice samples) in which only the numerals are given. In each instance, the subject is required to add the appropriate symbol. This test, also known as a substitution test, is regarded as requiring the association of symbols and involving speed and accuracy of performance. It involves, also, visual memory. The purely motor factor, it has been found, is relatively unimportant, except in the case of illiterate persons who are not accustomed to using pencil and paper.

FUNCTIONS INVOLVED IN THE SUBTESTS

The functions involved in each of the eleven subtests may be psychologically analyzed as shown below. They indicate the processes that are operative in most effective performances on each of the tests. This analysis should be distinguished from a *factorial* analysis, which is a *statistical* procedure whereby an attempt is made to consolidate and reduce the number of nonstatistically analyzed functions on the basis of communality, and perhaps to re-name them.

<i>Subtest</i>	<i>Functions</i>	<i>Influencing Factors</i>
Information	Long range retention Association and organization of experience	Cultural environment Interests
Comprehension	Reasoning with abstractions ⁷ Organization of knowledge	Cultural opportunities Response to reality situations

⁷ "Reasoning with abstractions" generally involves the processes of both analysis and synthesis, with the use of symbols—language and number. The testee must first analyze the relationships existing among the members, or parts, of the whole problem; then he must reorganize and interpret and, at times, create new wholes in order to reach the desired solution.

<i>Subtest</i>	<i>Functions</i>	<i>Influencing Factors</i>
Arithmetic	Reasoning with abstractions Concept formation Retention (of arithmetical processes)	Attention span Opportunity to acquire the fundamental arithmetical processes
Similarities	Analysis of relationships Verbal concept formation	A minimum of cultural opportunities
Vocabulary	Language development	Cultural opportunities
Digit Span	Immediate recall Auditory imagery Visual imagery at times	Attention span
Picture Arrangement	Visual perception of relationships (visual insight) Synthesis of nonverbal material	A minimum of cultural opportunity Visual acuity at times
Picture Completion	Visual perception: analysis Visual imagery	Environmental experience Visual acuity at times
Object Assembly	Visual perception: synthesis Visual-motor integration	Rate of motor activity Precision of motor activity
Block Design	Perception of form Visual perception: analysis Visual-motor integration	Rate of motor activity Minimum of color vision
Digit Symbol	Immediate rote recall Visual-motor integration Visual imagery	Rate of motor activity

While the ability to verbalize and to make abstractions is not necessarily involved in the five nonverbal subtests, nevertheless it has often been observed by examiners that this ability does facilitate and expedite one's performance. For example, on the Block-Design test, it is possible to analyze and formulate the color and form relationships

of each design before beginning to reproduce it. In the Picture Arrangement test, some subjects will attempt to discern and state the story told by the group of pictures before starting to place them in the correct sequence. It is important to recognize this psychological fact in evaluating and analyzing performance on tests that are primarily nonverbal: namely, *that even with such types of tests, ability to verbalize and abstract may be and at times is one of the psychological functions involved.*

NEED FOR AN ADULT SCALE

It will be recalled that Binet's own scales were not suited to use with adults; nor was the Stanford Revision of 1916. And while the 1937 Stanford-Binet is better standardized at the upper levels, including three "superior adult" levels, there is reason to question its adequacy in testing superior adults. There is also, as has already been explained, the difficulty and inconsistency of using the mental-age index with superior adults; that is, those having intelligence quotients above 100, who would therefore have to have a mental age rating above that of the average adult.

Intelligence testing of adults was begun on a large scale in 1917 with the establishment of a psychological division in the United States Army, in World War I. At that time, the Army Alpha (verbal) and Army Beta (nonverbal) tests were assembled, and with them about one and three-quarters million men were tested. This experience in large-scale testing provided the impetus for the development, after the war, of a number of other group tests for adults. But these tests did not prove to be adequate, with the exception of several which were designed for use with selected and limited groups of our population, such as candidates for admission to colleges. (These will be presented in a later chapter.) The Bellevue scale was, therefore, developed and has been offered as an *individual* test, standardized for ages ten to sixty.⁸

STANDARDIZATION

The first step in standardizing the Bellevue scale was to adopt a view of intelligence which should serve as the framework, so to

⁸ Although the age distribution of subjects used in standardizing the scale was from seven to seventy, the author gives the age range for use as ten to sixty.

speak, within which the test items would have to fit. The general-factor (*g*) theory was adopted, which requires that there should be significant intercorrelations between the several parts of the scale, and that the scale in its totality should provide a valid index of the individual's general ability.

Having accepted the theory of a general factor, the author of the Bellevue scale then had to determine which types of test materials should be included in the scale in order best to measure that factor. On the basis of past experience, it was decided that both verbal and nonverbal materials provide the most adequate and representative content, rather than either one alone. Having arrived at this view, the author and his collaborators proceeded to select the particular *types* of subtests which had been widely used by many other psychologists and which experience and experimentation had proved were valuable. In some instances, specific *items*, already in use, were included within appropriate subtests; in other instances, it was necessary to create new items for each of the subtests.

THE POPULATION SAMPLE

The next step was the selection of the population upon whom the scale should be standardized. The basis used was this: the sampling of adult population should be based upon the occupational distribution of the country's adults, as shown in the United States Census of 1930. The very numerous occupational subdivisions of the census were combined to make ten comprehensive categories (such as agriculture, manufacturing, clerical, etc.). From the records of many adults who had already been tested and whose scores were available, a standardization population was selected so that their occupational distribution should correspond reasonably well with that of the 1930 census. The final adult standardization group included 1081 literate individuals (all white) ranging in age from 17 to 70. After this adult population had been selected, their educational status (educational level completed) was compared with that of the U. S. population at large, the assumption being that a similar distribution of educational levels of the standardization group would provide further evidence of its representative character. In this respect, only moderate correspondence was found, for on the whole the educational level of the standardization group is higher than that of the general population, although the range in both groups is from college graduates to illiterates.

Since the Bellevue scale was to be standardized for some ages below

adult levels, it was necessary also to utilize a population of children. For the selection of a representative population of the younger ages (sixteen and less), the chosen criterion was the age-grade distribution of pupils in the public schools of New York City. About 1300 children were then tested in "representative and average" schools of New York City and about 200 in nearby communities in New York and New Jersey. From these 1500, a number were selected (plus a small group of mental defectives from an institution) so as to yield an age-grade distribution that would fairly well approximate that of the New York City school population. Thus, the final standardization population of the younger group included 670 white subjects, ranging in age from seven to sixteen years, and in grade placement from "ungraded" to twelve (plus continuation schools).

VALIDITY

The nature of the content of the scale having been decided upon and norms of performance for a wide range of ages having been determined, the scale was subjected to several analyses intended to determine its validity.

Intercorrelations of subtests. These coefficients are always necessary to provide data with regard to the presence or absence of a *g* factor. For this scale, appreciable correlations would be required, as one aspect of its validity. The author reports the following coefficients (each subtest correlated with every other subtest):—

Range of coefficients:

.15 (PE = .034), Object Assembly with Digit Span (N = 355; ages, 20–34)

.72 (PE = .026), Similarities with Comprehension (N = 355; ages, 20–34)

.37 to .72 is the range of the highest three-fourths of the intercorrelation coefficients (N = 590; ages, 20–49)

More significant are the coefficients found between scores on each subtest when correlated with total scores of *all other* parts of the scale.⁹ These coefficients were—

⁹ For example, Information scores would be correlated with the total scores of all remaining parts. If the scores on a given part—say, Information—were correlated with total scores of the test *including* the Information scores, it is obvious that the resulting coefficient would be in part *self-correlation*; consequently it would be spuriously high.

Range of coefficients:

- .41 (PE = .029), for Object Assembly (N = 355; ages, 20–34)
- .73 (PE = .02), for both Similarities and Block Design (N = 590; ages, 20–49)
- .60 to .72 is the range of the highest three-fourths of the coefficients (N = 590; ages, 20–49)

Additional evidence consistent with the presence of a general factor is provided by the following data showing the correlations between total scores on various combinations of subtests:

Ages of subjects, 20 to 34 years

Number of subjects, 355

- .83 (PE = .018), for total verbal subtest scores with total performance subtest scores
- .90 (PE = .014), for four subtests combining verbal and performance with a similar combination of four other subtests

The foregoing coefficients or intercorrelation cannot be said to be spuriously high, in spite of the fact that chronological age of the subjects varied considerably; for the subjects were adults, and in adult groups there is no continuous increase in score with successive ages, as there is in the case of children who are still developing in mental capacity. In calculating similar correlation coefficients for groups of children, it would be necessary to eliminate chronological age as a factor. If this were not done the coefficient would appear to be higher than it actually should be, since both variables, being correlated, would reflect the influence of a third factor: namely, chronological age. Similar statistical information obtained with a representative group of adolescents should have been provided in the process of standardization.

Correlation with schooling. The reader will recall that beginning with Binet, the amount of schooling or quality of educational achievement—or both—were used as criteria of validity. Hence, the ratings obtained on the Bellevue scale, for the adult standardization population, were correlated with number of years of schooling, the coefficient being .64. (With mental defectives omitted, $r = .53$.) This coefficient of .64 falls within the range of coefficients commonly found when these two variables are correlated.

Correlation with teachers' judgments. Teachers' estimates of pupils' intelligence, rated on a six-point scale, were used as a validating criterion. For a group of adolescents (74 in number) in a trade school, the correlation coefficient between Bellevue IQ's and teachers' ratings was .52; for another group (45 in number) in a general high school, the coefficient was .43. The number of cases in each of these instances was too small to be very meaningful.

Increase and decrease of scores with increasing age. Tables of the scale show rises of mean scores until the age of 22.5 years, and thereafter slow decline. The rise in mean scores to age 22.5 is quite inconsistent with norms of other tests. The reasons for this have yet to be definitely determined. Was the standardization adult population more adequate than those of other tests? Or was it a selected group? Are the test materials more adequate and better scaled? Is the increase in norms an artifact due to methods of scoring?

Constancy of means and standard deviation of IQ's at various ages. The means are reasonably constant, ranging from 98.75 at the age interval 17-19 years, to 101.25 at age 10, with about sixty percent of the means falling between 100 and 101. The standard deviations range from 13.2 IQ points at age 10 to 16.85 at the age interval 50-59, while about half the standard deviations are between 14 and 15 points. (See page 136 regarding significance of constancy of means and standard deviations of IQ's.) These data indicate reasonably good satisfaction of this criterion.¹⁰

Range and distribution of intelligence quotients. The range of 1508 cases, ages 10 to 60, as shown in graph form, is from about 45 to 145. The distribution curve for this group is slightly skewed in a negative direction;¹¹ that is, the IQ's fall within a somewhat narrower range above 100 than below. The author of the scale holds that the skewness does not bring into doubt the scale's validity; for he believes the widespread assumption of a symmetrical, bell-shaped distribution (Gaussian) is a "mistaken belief." Many psychologists, however, would regard his curve of distribution as being a satisfactory approximation

¹⁰ The coefficients of variation $\left(\frac{SD}{M} \times 100\right)$ for the IQ's range from 13.04 at age 10 to 15.67 at the age interval 30-34.

¹¹ Wechsler interprets his curve as being "considerably skewed."

to the symmetrical curve which Terman assumed as necessary in standardizing the Stanford-Binet scales, and which Binet himself implied as desirable in the construction of his own scales.

Known groups. In one study, two groups were differentiated on the basis of total scores: namely, (1) a borderline group, having IQ's between 66 and 79; and (2) a mentally defective group, having IQ's between 50 and 65. The problem was to determine whether each of the eleven subtests contributes significantly to the differentiation of the two groups. Since the mean scores on each of the subtests for the two groups did differentiate, and since the differences between the means were in the directions that should be expected, it was concluded that each subtest did contribute to the over-all differentiation, although Digit Span and Object Assembly contributed relatively little.¹²

A second study used only the verbal subtests with naval recruits as subjects. The problem here was to learn whether these subtests distinguish between (1) the mentally defective and the borderline, or (2) the borderline, the dull normal, and the normal. The findings indicated that each of the verbal subtests contributed to the differentiations between these groups. The Digit Span subtest in this instance, however, proved to be as effective as the others.¹³ This criterion of validity has still not been applied to the Bellevue scale with the same thoroughness as it has been applied to the Stanford-Binet.

Correlation with the Stanford-Binet. In validating a new scale designed to test intelligence, it is common practice to correlate results obtained by means of the new device with ratings obtained on the Stanford-Binet. To do this is in effect to accept the Stanford-Binet as a reasonably valid and reliable scale, and hence as one sound criterion of validity against which to evaluate the new scale.

Using seventy-five cases (ages 14-16), the author of the Bellevue scale reports that Bellevue IQ's and Stanford-Binet IQ's when corre-

¹² D. Wechsler, et al., "A Study of the Subtests of the Bellevue Intelligence Scale in Borderline and Mental Defective Cases," *American Journal of Mental Deficiency*, Vol. 45, pp. 555-558, 1941.

¹³ R. J. Lewinski, "Discriminative Value of the Subtests of the Bellevue Verbal Scale in the Examination of Naval Recruits," *Journal of General Psychology*, Vol. 31, pp. 95-99, 1944. Also H. M. MacPhee, et al., "The Performance of Mentally Subnormal Rural Southern Negroes on the Verbal Scale of the Bellevue Intelligence Examination," *Journal of Social Psychology*, Vol. 25, pp. 217-229, 1947.

lated yielded a coefficient of .82 ($PE = .026$). Six correlational studies of Bellevue and Stanford-Binet ratings, made by others, have yielded coefficients ranging from .57 ($PE = .04$) to .93 ($PE = .01$). Although coefficients of partial correlation, with chronological age constant, are not given, it is doubtful if the age range in any of these studies is such as to have raised appreciably the size of the coefficient. In the two instances where the coefficients were below .80, the subjects were college freshmen; hence the relatively lower coefficients (.57 and .62) may be attributable to either of the following conditions: the fact that the group is relatively homogeneous, with resultant constriction of range and reduction in correlation; the fact that the tests are not as reliable at the upper extreme, so that errors of measurement make high correlations unlikely.

Comparative studies, using the Stanford-Binet and the Bellevue, have not been as numerous as might have been expected in view of the wide use of both as clinical instruments. Unfortunately, too—but understandably since the Bellevue's clinical application has been emphasized—a large percentage of the comparative studies have used hospital patients and clinic referrals as their subjects, to the neglect of school pupils and others who fall within categories of "normal" behavior and adjustment.

The available data, however, do show that there is very substantial correlation between these two scales, particularly when the Bellevue Full Scale IQ's and Verbal Scale IQ's are correlated with Stanford-Binet IQ's. On the other hand, as would be expected, correlations between Stanford-Binet IQ's and Bellevue Performance Scale IQ's are only moderate. Table 27 shows the distribution of correlation coefficients found in a number of representative studies.¹⁴

Since correlation coefficients indicate *relative* agreement of paired scores, but not *absolute* differences between them, it is necessary, in comparing the Stanford-Binet and the Bellevue Scales, to know the extent of actual IQ differences existing between the correlated values. On the whole it has been found that the differences are not very large. In studies of retarded and mentally deficient persons (say, the lowest decile group), the Bellevue yields somewhat higher IQ's. At the upper

¹⁴ Correlational studies between the Bellevue and group scales have yielded varying results, the coefficients ranging from $.39 \pm .07$ (PE), for Thorndike's CAVD test to $.81 \pm .04$ (PE), for the Henmon-Nelson. Of six other coefficients, two are in the .70's, two in the .60's, and two in the .50's.

level of mental ability, however (say, the highest decile group), the Stanford-Binet yields somewhat higher IQ's.

Comparison of IQ's is complicated by the fact that age of testees is also a factor. Taking the population samples *as a whole* (rather than only the extreme groups), the following are the general findings. (1) Within the age range of approximately 10 to 19 years, the Stanford-Binet IQ's tend to be somewhat higher. (2) From age 19 to about 35,

TABLE 27

*Coefficients of Correlation between the Stanford-Binet
and the Bellevue Scales
(Frequencies)*

<i>r</i>	With Full Scale IQ	With Verbal IQ	With Performance IQ
.90	5	2	
.85	2	2	
.80	1	1	1
.75	2	1	
.70			1
.65			1
.60	1	1	
.55			1
.50			2
.35			1

the intelligence quotients tend to be about the same. (3) Above the age of 35, the Bellevue intelligence quotients tend to be somewhat higher.¹⁵

In the case of any given individual, therefore, comparison of Stanford-Binet and Bellevue IQ's must take into account both factors: ability level and chronological age. In a given instance, a person's age and ability level may be such as to increase or decrease the difference between the IQ's obtained with the two instruments.¹⁶

¹⁵ Number (3) is just what one should expect in view of the method employed in calculating Bellevue IQ's. The method is explained later in this chapter.

¹⁶ A definitive answer to the question of the comparability of the IQ's of the two scales will have to be based upon an investigation that is representative of the general population rather than heavily weighted with hospital and clinical subjects, as is now the case. Also, such an investigation must approach the problem in two ways: (1) by analyzing IQ's separately for each of the age-groups; and (2) by analyzing the IQ's separately at each of the ability levels for each of the age-groups.

On the whole, the correlation coefficients at hand, and other comparative data found between the two scales, are reasonably satisfactory and indicate that each has much in common with the other so far as concerns psychological functions being tested and ratings of ability levels.

Prognostic efficiency. The author of the Bellevue scale, as a final test of its validity, applies the pragmatic criterion. He states: "How do we know that our tests are 'good' measures of intelligence? The only honest reply we can make is that our experience has shown them to be so. If this seems to be a tenuous answer we need only remind the reader that it has been practical experience which has given (or denied) final validity to every other intelligence test. . . . Empirical judgments, here as elsewhere, play the role of ultimate arbiter. In any case, all evidence for the validity of a test, whether statistical or otherwise, is inevitably of an indirect sort and, in the end, cumulative rather than decisive."¹⁷ In other words, it has been found by the author of the scale and by others that it works with reasonable satisfaction in clinical practice.

In evaluating the prognostic effectiveness of the Bellevue or the Stanford-Binet scale, or any other, one must bear in mind the age range for which a particular instrument is most effective. A blanket statement about a scale's diagnostic effectiveness is not warranted.

Furthermore, conclusions regarding the prognostic efficiency of a psychological scale are dependent upon the soundness of the clinical diagnoses with which the scale's findings are compared. Herein lies the major problem and weakness in attempts to determine prognostic efficiency of a psychological scale; for the determination of diagnostic categories is difficult, often unreliable, and subject to the diagnostician's theoretical orientation. The major exception to this statement is the diagnosis of mental deficiency, for the determination of which a sound individual test of general intelligence is the most valid single instrument, when administered and interpreted by a qualified psychologist. For this purpose, both the Stanford-Binet and the Bellevue scales have proved to be most valuable. It must be added at once, of course, that a diagnosis of mental deficiency is usually not made upon the basis of IQ and MA alone, although in some cases findings with a single test are so clear and unequivocal as to suffice.

¹⁷ Wechsler, *op. cit.*, pp. 127-128.

RELIABILITY

It is a noteworthy fact that very little research has been done on the reliability of the Bellevue scale. On the other hand, a great volume of material has been published on its clinical uses and interpretation, due very probably to the fact that the scale was developed and originally used primarily by a hospital staff of psychologists, and due also to the fact that it has been used most extensively by clinicians since its publication. The neglect of reliability studies, while research emphasis has been placed upon studies dealing with differential diagnosis, intellectual deterioration, intellectual changes under treatment, etc., is regrettable; for the basic soundness of an instrument should be studied at least collaterally with its application.¹⁸

Wechsler's manual itself reports only very meager and inadequate data on reliability: namely, 52 individuals retested at intervals of one month to one year, with the results shown in Table 28.

TABLE 28
*Retest Correlation Coefficients for the
Bellevue Scale*¹⁹

Ages	N	Rho *	P.E.
10-13	32	.94	.013
20-34	20	.94	.018

*Rank order correlation coefficient

Since publication of the foregoing data, even the few available reliability studies have dealt almost exclusively with abnormal subjects, principally psychoneurotics and schizophrenics.²⁰ Table 29 shows the range of correlation coefficients found with such groups for each of the subtests and IQ scales.

¹⁸ The ready use of a new and promising clinical instrument is understandable and justifiable, since clinicians, confronted by immediate, pressing and persistent problems of living human beings, cannot wait until experimental research has subjected the instrument to thoroughgoing tests of reliability and validity. In using an instrument, however, it is essential that we give due consideration to its limitations and to unanswered or only partially answered questions about it.

¹⁹ From Wechsler, *op. cit.*, p. 133. (By permission.)

²⁰ Because of the instability of such persons, they are not the most suitable subjects to use for the study of the inherent stability of a measuring instrument.

Examination of Table 29 shows: (1) that there is considerable variation in reliability among the subtests and that, in general, their reliability is appreciably below that of the scale as a whole; (2) that in all but one of the reports (in which the subjects were schizophrenics and $r = .55$) full scale reliability appears to be reasonably satisfactory, considering the instability of the groups used. (Other r 's were: .87, .84, .84, .87, .89, .90.)

TABLE 29
Test-Retest Reliability of the Bellevue Scale Reported
for Abnormal Groups
(7 Studies)

Subtest	Range of Coefficients
Information	.56-.99
Comprehension	.12-.78
Digit Span	.59-.77
Arithmetic	.68-.87
Similarities	.38-.93
Vocabulary	.90-.93
Picture Arrangement	.49-.86
Picture Completion	.32-.89
Block Design	.65-.87
Object Assembly	.31-.79
Digit Symbol	.34-.91
Verbal IQ	.76-.91
Nonverbal IQ	.52-.94
Full Scale IQ	.55-.90

From published reports, it appears that only one investigation, using a fairly adequate sampling of individuals, has been devoted to reliability of the Bellevue when administered to "normal" subjects.²¹ The test-retest method was employed. The age range was 20 to approximately 50 years. One group of 60 subjects was retested after a one-week interval; another group of 60 persons after a four-week interval; a third group of 38 subjects after a six-month period. The major findings were the following.

²¹ G. F. Derner, M. Aborn, and A. H. Canter, "The Reliability of the Wechsler-Bellevue Subtests and Scales," *Journal of Consulting Psychology*, Vol. 14, 1950, pp. 172-179. See also W. B. Webb and H. DeHaan, "Wechsler-Bellevue Split-Half Reliabilities in Normals and Schizophrenics," *ibid.*, Vol. 15, pp. 68-71, 1951. The split-half method is an inappropriate method to use with most of the Bellevue scale.

The mean score for every subtest and for the total scale increased for all three groups.

Increases in scores tend to be somewhat smaller as the retest interval is increased.

The smallest average increase was 0.3 points in weighted score for comprehension retest (after four weeks).

The largest average increase was 2.8 points in weighted score for picture arrangement retest (after one week).

Largest average increases (2 or more points in weighted score) were found for picture arrangement and object assembly.

Smallest average increases (less than one point in weighted score) were found for information, comprehension, and similarities.

Average changes in IQ's were: verbal scale, 4.4 points; nonverbal IQ, 9.1 points; full-scale IQ, 7.6 points.

Retest correlations and standard errors of measurement ²² for all subtests and the three IQ's are shown in Table 30. It will be noted

TABLE 30
Test-Retest Correlations and Standard Errors of
Measurement for the Bellevue Scale ²³
(N = 158)

Subtests	Correlations	S.E. meas.
Information	.86	.68
Comprehension	.74	1.21
Digit Span	.67	1.68
Arithmetic	.62	2.06
Similarities	.71	1.22
Vocabulary	.88	.73
Picture Arrangement	.64	1.82
Picture Completion	.83	.95
Block Design	.84	1.10
Object Assembly	.69	1.31
Digit Symbol	.80	1.06
Verbal IQ	.84	3.96
Nonverbal IQ	.86	4.49
Full Scale IQ	.90	3.29

²² For interpretation of standard error of measurement, see Chapter 1, p. 17.

²³ From G. F. Derner, et al., *op. cit.*

that, for the subtests, four of the coefficients are in the .60's (very low reliability); two are in the .70's (low reliability); five are in the .80's (satisfactory reliability for a subtest).

On the whole, it appears from this study, using subjects within the "normal" range of personality and behavior, that the reliabilities of the Bellevue scale are not high enough in six of the eleven subtests to warrant their use independently. This is a significant finding, especially as it relates to the use of the subtest profile for diagnostic purposes (discussed in Chapter 15). The reliability coefficients for the Verbal and Nonverbal IQ ratings are reasonably satisfactory, though not as high as some psychologists would demand. The full scale IQ reliability, however, is quite satisfactory. The standard errors of measurement, it will be noted, conform rather closely with the correlations. The general conclusion, then, is that while the subtests individually do not show a high degree of reliability, *the scale as a whole yields reasonably reliable results*, as indicated by the correlation coefficients and the standard errors of measurement.

These results indicate, again, the significance and value of measuring a number of functions; for not only does such a measure yield a more representative index, but there is a greater probability that daily or periodic fluctuations in performance will be compensated for through the testing of a number of representative functions.

SCORING AND IQ CALCULATION

Scoring. All parts of this scale are scored on a point basis. For some subtests, the earned raw score is simply the number correct, each item being scored either plus or minus (e.g., information). Or, as in the case of comprehension and similarities, the score for each item is 0, 1, or 2, depending upon quality of the response. In other parts, as in arithmetical reasoning and block design, the earned raw score is based not only upon correct responses but upon the time taken to solve the problem. Thus, the factor of speed of performance is involved in sections of this scale, especially in nonverbal subtests.

The *raw score* for each subtest is first obtained by simple addition of the credits on the items in that part. This raw score is then converted into a weighted score (a type of standard score), by means of a conversion table. The purpose of this conversion is the customary one of placing all subtest scores on a comparable basis. The weighted

scores for all parts of the scale are then added to obtain the full score upon which the "full scale IQ" is based. Also, the weighted scores of only the six verbal parts are added to get the verbal score, upon which the "verbal scale IQ" is based. Similarly the weighted scores of the five performance tests are added to get the performance score and "performance scale IQ."

The following well-known formula was used for equating each subtest's raw score into *weighted scores*.

$$X_2 = M_2 + \frac{SD_2}{SD_1} (X_1 - M_1),$$

in which

M_2 = an arbitrarily assigned mean (10)

SD_2 = an arbitrarily assigned standard deviation (3)

X_2 = the weighted score to be found

M_1 = the mean of the subtest's raw score

X_1 = the particular raw score to be converted to a weighted score

What this formula does is: (1) assign an arbitrary and *uniform mean score* to all subtests; (2) multiply each individual score's deviation from its mean by a *constant ratio*; (3) add the result to or subtract it from the assigned mean.²⁴ By using this formula, scores are so converted that each individual maintains his relative status on each subtest. And in the case of any given person's subtest scores, differences between scores will be attributable, theoretically, to differences in his performance level rather than to differences in the weighting of each subtest in the total. It is thus possible to vary the number of items in each of the several subtests without giving any of them too little or too much weight in the total score upon which the IQ is based.

On the Bellevue scale, the reason for converting raw scores into weighted scores is this: the possible maximum raw scores vary in the several subtests of the scale. If, therefore, the raw scores were simply

²⁴ The best way for the student to see how this formula works is to substitute several sets of values in it and to observe the outcome. The logic of the process will then be more readily apparent. For example, if the two following sets of values are substituted in the formula, the process will be clear. Assume the following data for one subtest: mean score = 12, SD = 4, $X = 15$; while for a second subtest the corresponding values are 24, 8, and 30. The results will show the same weighted score for both subtests (12.25) because the individual's relative status on each subtest was identical with that on the other subtest, even though the raw scores differ.

added to obtain an individual's rating on the scale, each of the parts would carry a different weight in the total; each part would have the possibility of contributing differently to the final result—some more heavily than others. The raw-score units of one part of the scale would not have the same significance as those of other parts. If this were the case, then implicit in the scoring would be the assumption that certain of the psychological functions being tested should be regarded as more important than others in the total score and in getting an index of intelligence. The Bellevue scale, however, is scored on the principle that all the functions tested are equally important; hence, the part-scores should be equally weighted so that each part may contribute as much to the total as any other part.

Calculating the IQ. After the verbal score, performance score, and full score have been obtained (in converted units), the three corresponding intelligence quotients are found in tables prepared for that purpose. In calculating IQ's for the Bellevue scale, the *basic principle* used differs from that in the Stanford-Binet and in most other scales. The *principle* employed is that an individual's intelligence quotient should be determined by the relative extent to which his weighted score (full, verbal, or nonverbal) deviates from the mean weighted score of *his own age group*. The detailed method of actually determining Bellevue intelligence quotients involves several steps and assumptions.

First, the mean weighted score and the standard deviation for each age level are calculated. Second, the weighted scores at each age level are converted into standard scores (*Z* scores. See Chapter 2, page 45). Then, third, it is assumed that a value of .6745 of a standard score shall be equated with an IQ of 90.²⁵ Hence, this assumption means that the probable error (.6745 SD) of Bellevue IQ's is arbitrarily set at 10 points; and the standard deviation of IQ's at approximately 15 points, since the PE equals .6745 of the SD.²⁶ In other words, according to this device, in a "normal," symmetrical distribu-

²⁵ The reason why .6745 *Z* is taken is this: the index called "probable error" (PE) is .6745 of a standard deviation. The standard score, it will be recalled, is an index given in terms of the standard deviation. The PE, therefore, is convenient and easily calculated when the standard scores are known.

²⁶ This device makes Bellevue IQ's approximate the most probable SD of the Stanford-Binet distribution of intelligence quotients: namely, about 16 points.

tion of Bellevue IQ's, fifty percent of the IQ values will fall between 90 and 110 (± 1 PE), since the PE sets the limits of the middle fifty percent of the scores of a distribution; and about two-thirds (68.26 percent) of the values will fall between 85 and 115 (± 1 SD).

On the basis of the foregoing assumption, it is possible to assign an IQ value to any weighted standard score. In effect, what the procedure amounts to is that the standard score is converted into a probable error value; this value is multiplied by 10; and the result is added to or subtracted from 100, depending upon whether the standard score is plus or minus. The calculations are shortened and facilitated by a formula for the purpose.²⁷ The intelligence quotients found in this manner should be called "deviation IQ's."

Criticism of the Bellevue Method. When this method of determining intelligence quotients is used, the following points should be noted. This method consistently compares (by means of IQ's) an individual's test score with the mean for his own chronological age group up to 60 years, whereas on the Stanford-Binet and other scales a person whose CA is greater than the age at which "average adult MA" is reached is compared with a given and fixed maximum age level. Thus it will be recalled that on the 1916 Stanford-Binet, in calculating the IQ of a person 16 years of age or older, the maximum value in the denominator is 16 (MA/16) regardless of his actual age. When using the 1937 Stanford-Binet, the maximum denominator is 15. In other words, the method used with the 1937 Stanford-Binet relates the test performance of a person above age 15 to the performance level of an average group at a specified maximum age (15); the Bellevue method, at all ages up to sixty, relates a person's test performance to the average performance of *his own* age group. Also, then, the Bellevue method, for the purpose of IQ calculation, obviates the necessity of determining age of "average adult MA"—always a difficult problem and as yet uncertain.

One weakness of the Bellevue method of calculating IQ is this: the same, or constant, objective performance on the test (i.e., the score) will rate an individual higher with increased age after the maximum, since his constant score will be compared with steadily declining age norms. For example, a person who earns an IQ of 60 at age 20, main-

²⁷ See Wechsler, *op. cit.*, pp. 219-220.

taining a constant score to age 50, would have an IQ of 78 at age 50. To maintain an IQ of 60 there would need to be an average rate of decline consistent with an IQ of 60.

Another point concerns the question of when an individual's mental level, *as measured by the tests*, begins to decline. This, too, presents a question as yet not definitively answered. There is also the question of the *rate* of decline in functions being tested during each of the several stages of adulthood. The Bellevue IQ's would not provide any evidence toward answers to either of these questions. But regardless of specific answers to these two questions, data show that with the Bellevue scale there is some slight decline in average *test scores* between the ages of approximately 25 and 45, and that rate of decline in average test scores increases thereafter.²⁸ Obviously, therefore, after the period of decline (however moderate) sets in, as shown by the tests, an individual's IQ will begin gradually to decline if the Stanford-Binet method is used; whereas if the Bellevue method of calculating IQ is used, an individual's rating will decline only if he loses ground *with reference to the average of his own age group*, rather than with reference to a more or less hypothetical "average adult level." If his losses are at the same general rate as those of his age group, his Bellevue IQ will remain relatively constant. If, however, his losses are less than the general rate, his Bellevue IQ will rise.

SPECIAL FEATURES OF THE BELLEVUE SCALE

The Bellevue scale provides a scheme for calculating a "deterioration quotient" based on the premise that certain types of tested mental processes decline more rapidly than do other types, and that the difference between rates of decline, as between these two types, in the case of any given person indicates his relative degree of deterioration. In other words, there are certain tested functions that hold up with age and others that do not hold up with age. This index

²⁸ The reader should note that we have emphasized decline in "test score." This does not necessarily mean that on the whole a person becomes progressively "less intelligent" even before the effects of senescence become apparent. While it is true that there is some loss in average test scores after about age 25, it is also true that some mental traits, as yet unmeasured by intelligence tests, increase in effectiveness through an extended period of adulthood and more than compensate for losses in the processes measured by current scales. This view is borne out by the facts regarding ages of maximum achievement of scholars, scientists, writers, and artists.

will be dealt with in more detail in a later chapter, together with other tests devised to measure deterioration of mental abilities.

A second feature of the Bellevue scale is its emphasis upon "scatter analysis"—that is, analysis of an individual's performance on the several parts of the scale for the purpose of facilitating clinical analysis of the subject's performance. Such analysis may lead to diagnostic inferences concerning personality characteristics and behavior disorders due to organic brain disease, psychosis, psychoneurosis, adolescent psychopathy, and mental deficiency. Here again, this application of the scale and clinical evidence, supporting and contrary, will be presented in a subsequent chapter on clinical uses and interpretations of tests.

CRITICISMS AND EVALUATIONS

The Bellevue scale is being quite widely used in the measurement of adult intelligence, especially in psychological clinics. Implicit in the wide use of a testing instrument is endorsement and acceptance of the scale, at least as one of the more satisfactory of those available at the time.

Was the population sample adequate? Since the 1751 persons upon whom this scale was standardized were from New York City and nearby communities, the adequacy of the population sample may be and has been seriously questioned. It would be highly desirable for the author of this scale, and his associates, to assemble and publish a frequency distribution of scores and intelligence quotients obtained with "normal" groups in various sections of the country. These data should be separately presented for each of the several age levels; they should be analyzed for socio-economic differences and sex differences, to determine if these are significant.

The occupational distribution of the standardization population was a moderate approximation to the 1930 census which was taken as the basis for the selection of the population sample. Furthermore, the validity of occupational regrouping in the process of standardization may be questioned.²⁹ Also, in the standardization population, there were more persons from the upper educational levels than there were in the general population.

²⁹ See Wechsler, *op. cit.*, p. 111.

In view of the foregoing considerations, the norms of the subtests and the totals should not be regarded as final.

Are the subtests a variety of disconnected types of tests? With the exceptions of object assembly and, to a lesser extent, picture arrangement, the intercorrelations of the subtests are all significant and marked. This would suggest that the Bellevue subtests are measuring one or more common factors to an appreciable extent. The *intercorrelations* between verbal and nonverbal subtests are not, on the whole, as high as *intracorrelations* within each of these categories; but with the exceptions noted the *intercorrelations* are indicative of a common factor or common factors.

This scale has not been adequately analyzed factorially. One analysis, however, finds that a "first factor" (general factor) accounts for from 27 percent to 50 percent of individual differences in scores, the weight of the general factor varying apparently at different age-levels.³⁰ The high correlations between the Stanford-Binet and the Bellevue scales indicate that the two instruments have much in common as regards the psychological functions measured by them. And, it will be recalled, the Stanford-Binet has been found to measure primarily a general factor. The question of factors as determined from analysis of Bellevue scale results themselves, is, however, one that should be subjected to further comprehensive investigation.³¹

Are the verbal subtests culturally unfair to some persons? The answer to this question is the same as the one given for the Stanford-Binet. In addition, it may be said that in this scale, the verbal materials in Comprehension, Similarities, and Arithmetic are stated in such terms as place very little premium upon educational or other

³⁰ I. Lorge, *Third Mental Measurement Yearbook*, O. K. Buros, editor, New Brunswick, N. J.: Rutgers University Press, 1949, p. 393. See also B. Balinsky, *An Analysis of the Mental Factors of Various Age Groups from Nine to Sixty*, Genetic Psychology Monographs, Vol. 23, pp. 191-234, 1941.

³¹ The need of such research is still present even though a few factorial studies have been published recently. These studies are, unfortunately, limited to atypical groups. See J. Cohen, "A Factor-Analytically Based Rationale for the Wechsler-Bellevue," *Journal of Consulting Psychology*, Vol. 16, pp. 272-277, 1952. This report deals with psychoneurotics, schizophrenics, and brain-damaged individuals. The findings, therefore, are not as universally applicable as the title implies. Also: J. E. Birren, "A Factorial Analysis of the Wechsler-Bellevue Scale Given to an Elderly Population," *ibid.*, Vol. 16, pp. 399-405, 1952. Since the subjects of this report were between 60 and 74 years of age, the findings can hardly be regarded as representative.

cultural advantages. Like all tests of information and vocabulary, performance on these in the Bellevue is dependent, in part, upon opportunity to learn, whether in school, home, or through the intellectual exploitation of all aspects of one's environment.

Are some of the test items obsolete? Items in any and all tests must be reviewed periodically for possible obsolescence. In the case of the Bellevue, a number of the items, especially in comprehension, information, vocabulary, picture arrangement, and picture completion should be re-examined and re-evaluated. Also for a number of items the satisfactory, partially satisfactory, and unsatisfactory responses should be reviewed and revised in the light of responses that have been obtained since the scale's original publication.

Is the factor of speed important in the Bellevue scale? Unlike the Stanford-Binet, in which very few test items are timed, Bellevue scores are significantly affected by the speed factor. Speed of performance yields additional credits in the following subtests: arithmetic, picture arrangement, object assembly, digit symbol, and block design. Thus, in the total score, speed of work is combined with power (or ability level). Although in general speed and power are highly correlated, it is also a fact that response time slows down with age. Thus, since the Bellevue scale is designed for adults, to an important degree it measures, especially in later adult years, decline in speed of response and not necessarily decline in power. This factor must be kept in mind when, in a later chapter, we consider the "decline" of abilities and the suggested "deterioration" index.

Do the nonverbal subtests involve visual acuity? Although no experimental data are available in answer to this question, not a few users of the Bellevue scale have observed that visual acuity might be a factor in some instances. The subtests most likely to make some demands upon visual acuity are picture arrangement and picture completion. And, of course, color blindness must be eliminated as a factor in the block design subtest.

Are the reliability coefficients satisfactory? Available data indicate that the total verbal scores and the total performance scores have a fairly satisfactory degree of reliability; while the full scale scores have a degree high enough to satisfy the standards of most psychologists. The reliabilities of the subtests, however, are not high enough, when

taken individually, to be used uncritically for differential diagnosis or for evaluating deterioration of mental ability, since accuracy in both of these matters depends upon accuracy of retest results. More research is necessary on the several reliabilities for a normal population rather than for hospital and clinical cases.

Should the Bellevue scale use the IQ? There is warranted criticism of the use of the term "intelligence quotient" for the index derived with this scale, because the formula employed really changes the conception of the IQ and thereby confuses the meaning of the term. Meanings of scientific terms are established by priority and usage; and usage had established a conception of the IQ as developed by the Binet revisions and by group tests which followed the same basic principle as regards the maximum denominator in the formula ($IQ = MA/CA$). Since the IQ has been and is being used with the Bellevue scale, it is necessary to bear in mind that it is a "deviation IQ."

Should the verbal and the performance scores be combined? The tables of norms for this test show that the maximum verbal score norm is reached at the age of 22.5 years, while the maximum performance score norm is attained at the 16.5–18 years. Maximum full scale norm is found at the age of 22.5 years. In view of these differences, we may question the wisdom of combining verbal and performance scores, particularly after the age of 18, when one group of functions (performance) no longer develops differentially, while the other group of functions (verbal) does so develop. It is quite possible that these two sets of tests and functions are sufficiently different after the age of 18 so that they should be separately scored. It is also possible that the value of the scale is reduced by combining the two sets of subtests. Furthermore, it is possible that some of the inadequacies of the Bellevue Scale findings may be attributable to combining the two types of subtests.

Is the Bellevue scale clinically useful? Judging from its widespread use in clinics and hospitals and from the empirical judgments of many clinicians, it appears that this scale has been of considerable value. Used with scientific judgment and with knowledge of its limitations and its tentativeness in some aspects, the Bellevue scale can be a very useful instrument for estimating intelligence of adolescents and adults. An important qualification must be added: namely, that this

scale is not adequate to measure and differentiate the highest levels of ability; at least the upper five percent of the population.

The Bellevue scale is a valuable addition to other testing and diagnostic devices, such as the Stanford-Binet, the Arthur Performance Scale, the Babcock test, and others which will be presented. Between most of these scales there are significant correlations; the exclusive clinical effectiveness of one or the other has yet to be established. For the present, and no doubt in the future, clinicians (in schools, hospitals, and elsewhere) will use a given scale or a combination of scales as occasion demands and as their clinical insights suggest.

Some psychologists give considerable weight to the fact that the Bellevue is so constructed that it is possible to analyze an individual's scores in terms of his variations (consistency or inconsistency) on the several parts of the scale, especially since attempts have been made by the author of the scale and by others to specify the psychological functions being tested by each of the several parts.

The validity of the Bellevue in identifying personality and behavior disorders has yet to be unequivocally demonstrated. In spite of the fact that this is the area in which most of the evaluative studies of the scale have been made, clinical findings are by no means definitive.

In attempting to diagnose personality and behavior disorders on the basis of the pattern or profile of scores on the Bellevue or other scales, it must be remembered also that different educational backgrounds and cultural factors, quite unrelated to personality and behavior disorders, could account, to some degree, for an individual's inconsistency of performance on the several parts of the scale. It has been found, too, that individual variations in interests, as distinguished from personality disorders, find expression in different patterns of mental activities and are reflected in subtest variations. However, the results obtained by means of the Bellevue, plus clinical experience and acumen, provide a valuable combination for study of individual differences and individual mental functioning.

General comment. Since its appearance in 1939, the Bellevue scale has been widely reviewed and evaluated.³² Judgments have varied

³² See O. K. Buros, ed., *op. cit.*, pp. 386-398; and Buros, *The Fourth Mental Measurement Yearbook*, Highland Park, N. J.: The Gryphon Press, 1953, pp. 473-476.

from enthusiastically uncritical acceptance to destructively critical rejection. Some critics have concentrated, and justifiably, upon unwarranted assumptions and statistical inadequacies, to the exclusion of other and positive aspects. At the other extreme are the critics who have ignored the scale's defects and limitations while concentrating upon and lauding its practical value. Most evaluations, however, have been moderate in that they have pointed out the contributions, values, and possibilities of the scale, while clearly indicating its defects and doubtful assumptions. Experimental research and competent practical application can proceed simultaneously, each facilitating the other.³³

THE 1955 REVISION OF THE BELLEVUE SCALE

Early in 1955 a revised edition of this scale is scheduled for publication. This new edition, it appears, will meet some of the adverse criticisms directed against the original scale. The revised version (to be known as the Wechsler Adult Intelligence Scale) does not introduce any new principles in its content, construction, organization, scoring, or IQ derivation. The main changes are in the revision of some content, extension of the population sample, and in improvement in directions for administering and scoring. Some of the major revisions (supplied by The Psychological Corporation prior to publication of the scale) are the following.

Content. Range of difficulty has been extended, chiefly downward in order to assure a score for the lower level of mentally deficient subjects. Upward extension in difficulty has been slight. Progression of difficulty from item to item has been improved. Obsolete items have been replaced. Items having poor "item validity" and those overlapping others in content have been replaced, as have those that were ambiguous. Illustrations in the picture completion subtest have been

³³ Some of the other questions asked in connection with the Stanford-Binet might appropriately be asked about the Bellevue. But since the principles involved and the replies would be the same, they have not been repeated.

Form II of the Bellevue scale has been made available (New York: The Psychological Corporation, 1946). This form will not be discussed because: (1) it is, presumably, identical with Form I in respect to underlying principles and types of test materials; (2) the standardization information provided is too limited, and the data provided indicate that Form II does not meet the criteria necessary if it is to be used as an alternate scale: that is, equal or very nearly equal means, deviations, and distributions obtained with the same population sample.

more clearly drawn. The vocabulary subtest has been revised so as to produce a fairly normal distribution of scores for a representative sample of the population. Maximum scores in verbal and nonverbal subtests and in the full scale are reached by the 25–34 year age-group.

Population sample. Norms are based upon a sample of 1700 persons, 850 of each sex, selected from four major geographic areas. The subjects ranged in age from 16 to 64 years. The age range was divided into seven age-groups, within each of which the numbers were proportioned according to the 1950 U. S. census with respect to geographic area, race (white and nonwhite), occupation, urban-rural, and years of formal education. Supplementary data were also obtained for a sample of older persons ($N = 352$) above 65 years of age.

Reliability and Validity. For the separate subtests, the reliability estimates range from .66 (picture arrangement) to .96 (vocabulary). Reliabilities for the three IQ's are: verbal scale, .96; performance scale, .93; full scale, .97. Intercorrelations among the eleven subtests range from .30 to .85, while the correlation coefficient for total verbal scores vs. total performance scores (ages 18–24) is .77.

THE WECHSLER INTELLIGENCE SCALE FOR CHILDREN (1949)³⁴

Description. This scale for children from five through fifteen years of age is built on the same principles and in the same form as the Bellevue scale for adolescents and adults: verbal subtests, performance subtests, a verbal IQ, a performance IQ, and a full scale IQ.

The subtest types are identical with those of the older scale, with the following exceptions: digit span is made optional; an optional maze test has been added; and in place of digit-symbol, a coding test has been substituted, in which various lines in varied positions (single, double, circle) are associated with geometric figures (star, circle, triangle, cross, rectangle).

Standardization Population. The scale was standardized on a sample of 100 boys and 100 girls at each of the eleven age levels, each child being tested within one and one-half months of his mid-year. In ef-

³⁴ New York, The Psychological Corporation.

fect this means that children were selected at the half-way mark between birthdays; and half-way was defined as being between four and a half months and seven and a half months (excepting the feeble-minded, nearly all of whom were within two months of their mid-year).

Selection of the 2200 children was based upon: (1) rural-urban residence; (2) father's occupation; and (3) geographic area. The proportions in these sampling factors were based upon U. S. Census data for 1940, ". . . with some adjustment for the shift of population toward the West." In the final selection of the standardization sample, geographic area percentages are reasonably well satisfied; urban-rural percentages, less well; and father's occupation percentages, moderately.³⁵ On the whole, the standardization group satisfies the principles of sampling better than did the population sample used for the Bellevue adult scale. Still, 2200 children distributed among eleven age groups and over four very wide geographic areas (New England and Middle Atlantic States; North Central States; South Atlantic and South Central States; Mountain and Pacific States) are but a small handful. Extensive experimental use of this scale will be necessary to determine the adequacy of the norms based upon the population sample used in standardization.

Reliability Data.³⁶ Reliability coefficients were found for three age groups ($7\frac{1}{2}$, $10\frac{1}{2}$, $13\frac{1}{2}$), the number being 200 in each. The findings are summarized in Tables 31 and 32. It will be noted, from these data, that the subtest reliability coefficients vary markedly and are, on the whole, only moderate in magnitude. The IQ reliabilities, however, being from .86 to .96, fall within the range that is generally acceptable. These data demonstrate again the necessity of distinguishing between reliability of part of a scale and that of the whole scale.

³⁵ For detailed standardization, see Wechsler, *op. cit.*; and H. Seashore, et al., "The Standardization of the Wechsler Intelligence Scale for Children," *Journal of Clinical Psychology*, Vol. 14, pp. 99-110, 1950.

³⁶ The split-half technique was used to calculate reliability, except in the case of coding and digit span. For the former, results of coding tests A and B were used, since this is essentially a speed test. For digit span, scores on digits forward were correlated with scores on digits backward—a very questionable procedure since the two do not involve identical processes. Also, some question may be raised regarding the appropriateness of the split-half method for some of the other subtests. The test-retest method is much to be preferred for a scale of this type.

The standard error of measurement indicates the range of score within which the chances are approximately two to one that a subject's "true" score will fall in that particular subtest. Thus, the standard error of 1.20 for 7½-year-olds on picture arrangement indicates that the probabilities are two to one that an individual's "true" score on this subtest is within 1.20 points of his obtained weighted score. Likewise the standard error of 4.25 IQ points (full scale) for 7½-year-olds indicates that the probabilities are about two to one that an

TABLE 31

*Reliability Data: Intelligence Scale for Children*³⁷

Subtest

Age Group	Range of <i>r</i> 's	Mean	High <i>r</i>	Low <i>r</i>
7½	.59-.84	.67	Block Design	Comprehension and Picture Completion
10½	.59-.91	.76	Vocabulary	Digit Span
13½	.50-.90	.75	Vocabulary	Digit Span

IQ Reliabilities

	Verbal	Nonverbal	Full
7½	.88	.86	.92
10½	.96	.89	.95
13½	.96	.90	.94

(Digit Span, Coding, and Mazes are not included.)

individual's "true" IQ on this scale is within 4.25 points of his obtained IQ.

Conclusions on Reliability. The reliability coefficients and the standard errors of measurement must be taken into account when scores on the individual subtests are being interpreted or when differences in scores between subtests are being evaluated. The lower the reliability and the larger the standard error, the less is the confidence to be placed in judgments based upon scores of that particular subtest.

Since, on the basis of the standardization data, the reliabilities of the several IQ's are at a satisfactory level, it appears that considerably more confidence can be placed in those indexes than in the scores of the individual subtests (with the exception of vocabulary).

³⁷ From the *Manual*, p. 13. The Psychological Corporation. (By permission.)

Since there are marked differences between reliability coefficients of the subtests for each of the three age groups reported in the standardization data, it is highly desirable that separate reliability studies be made for each of the eleven age groups separately, especially at the extremes of the age distribution (5 to 15) for which the scale is intended.

TABLE 32
Standard Errors of Measurement: Intelligence Scale
for Children³⁸

Age Group	Range *	Mean	Subtest	
			High	Low
7½	1.20-2.45	1.74	Digit Span	Picture Arrange- ment
10½	.90-1.92	1.44	Digit Span	Vocabulary
13½	.95-2.12	1.47	Digit Span	Vocabulary
IQ Standard Errors **				
		Verbal	Nonverbal	Full
7½		5.19	5.61	4.25
10½		3.00	4.98	3.36
13½		3.00	4.74	3.68

* Standard errors of measurement of subtests are given in units of the weighted scores.

** Standard errors of intelligence quotients are given, of course, in IQ points.

Validity. *Subtest Intercorrelations.* In the manual for this scale, there are no data on the problem of validity as such. There are data on intercorrelations of the subtests. The assumption is that significant intercorrelations between subtests would validate the hypothesis that they and the scale as a whole measure common factors. However, the intercorrelation coefficients among the individual subtests are, on the whole, not as high as would be expected. At the 7½-year level, these coefficients are concentrated within the .20's and .30's; at the 10½ year level, they are concentrated within the .30's and .40's; while at the 13½ year level, they are distributed within the .20's, .30's, and .40's.

On the other hand *each verbal subtest* correlates quite significantly with *total verbal score*, the range for the three age groups being from

³⁸ From the *Manual*, p. 13, The Psychological Corporation. (By permission.)

.44 to .82, with the coefficients fairly evenly distributed over this range. The *nonverbal subtests* correlate somewhat lower with *total performance scores*, the range being from .32 to .68, with some concentration in the .50's.

The correlation coefficients between *total verbal scores* and *total performance scores* are, respectively, .60, .68, and .56 for these same age groups.

TABLE 33
Correlations between the Intelligence Scale for
Children and Other Scales
(5 Studies)

Other Scale	Subjects	Number	r
Arthur Point Scale	mentally defective	40	.79 (Full Scale)
“ “ “	“ “	40	.83 (Nonverbal Scale)
“ “ “	“ “	40	.47 (Verbal Scale)
Stanford-Binet, L	“ “	40	.76 (Full Scale)
“ “ “	“ “	40	.64 (Nonverbal Scale)
“ “ “	“ “	40	.75 (Verbal Scale)
Stanford-Binet	subnormals	70	.68 (Full Scale)
“ “ “	“	70	.69 (Verbal Scale)
Stanford-Binet	normals	49-53	.85 (Full Scale)
“ “ “	“	49-53	.82 (Verbal Scale)
“ “ “	“	49-53	.80 (Nonverbal Scale)
Arthur Point Scale	“	49-53	.80 (Full Scale)
“ “ “	“	49-53	.77 (Verbal Scale)
“ “ “	“	49-53	.81 (Nonverbal Scale)
Stanford-Binet, L	“	54	.80 (Full Scale)
“ “ “	“	54	.71 (Verbal Scale)
“ “ “	“	54	.63 (Nonverbal Scale)
“ “ “	“	332	.82 (Full Scale)
“ “ “	“	332	.74 (Verbal Scale)
“ “ “	“	332	.64 (Nonverbal Scale)

These findings indicate that, on the whole, while each subtest has only a very moderate amount of communality with the others taken singly, verbal subtests *combined* have much more communality with each individual verbal subtest.³⁹ The same is true of *combined* performance and separate performance scores.

Finally, the data indicate that *all the verbal subtests taken as a whole* have considerable communality with *all the performance sub-*

³⁹ Corrections are made in order to eliminate self-correlation.

tests as a whole. Yet, since the aforementioned coefficients of .60, .68, and .56 are fairly distant from unity, the measured abilities in one group (verbal) can be used only for a general approximation of abilities measured by the other group of subtests (nonverbal), and *vice versa*. The reporting, therefore, of verbal, nonverbal, and full

TABLE 34
IQ's of Intelligence Scale for Children Compared with
Two Other Scales
Means and Standard Deviations
(5 Studies)

WISC	Arthur Point Scale	S-B	Subjects	N
60(S.D.6) Full 65(S.D.13) Verbal 58(S.D.10) Perform. } 66(S.D.9) Full 67(S.D.7) Verbal 72(S.D.11) Perform. }	65(S.D.12)	56(S.D.5)	Deficients	40
		65(S.D.7)	Subnormal	70
100(S.D.15) Full 99(S.D.14) Verbal 101(S.D.15) Perform. }	95(S.D.16)	105(S.D.15)	Normal	49-53
102(S.D.11) Full 101(S.D.12) Verbal 104(S.D.11) Perform. }		106(S.D.11)	Normal	54
101(S.D.13) Full 103(S.D.14) Verbal 98(S.D.15) Perform. }		108(S.D.16)	Normal	332

scale IQ's with this instrument is a desirable, in fact a necessary, practice.

Correlations with Other Scales. Since the appearance of this scale, several reports have been published that deal with the correlations and IQ differences found between it, the S-B and the Arthur. The summarized data are given in Tables 33 and 34. Unfortunately, the findings of these studies must, with one exception, be regarded as only suggestive and quite tentative; for the number of cases in each is very small, or the coefficients have been affected by the age range of the testees.

The exception reports on 332 cases between the ages of 5 and 15.

The data given in Tables 33 and 34 are for the entire group. At the different ages, the r 's between Stanford-Binet and full scale IQ's vary from .75 to .90; for the verbal scale, between .65 and .90; for the performance scale, between .50 and .75. The table giving mean intelligence quotients and standard deviations indicates that the Wechsler scale tends to rate subnormal subjects somewhat but not markedly higher than does the Stanford-Binet. At the average level the reverse is true. The differences between the means are fairly marked in the study of 332 individuals. This being the most comprehensive and de-

TABLE 35
*Correlations of Intelligence Scale for Children with
School Achievement*

Scale	N	Range of r 's for Separate Subjects	Total Achievement Test
Full	54	.45-.71	.76
Verbal	54	.48-.60	.62
Nonverbal	54	.41-.64	.65
Full	18-21	.44-.81	—
Verbal	18-21	.47-.74	—
Nonverbal	18-21	.29-.74	—

tailed report of those herein reported, its findings carry the greatest weight.⁴⁰

On the basis of the research thus far reported, it is reasonable to conclude that full scale intelligence quotients and verbal scale intelligence quotients, on the one hand, and Stanford-Binet IQ's, on the other, have considerable communality of psychological functions being measured. The performance scale intelligence quotients have much less in common with the Stanford-Binet.

Predictive Efficiency. Validity of a test for children should also be evaluated in terms of its predictive efficiency with respect to educational achievement. In this area, too, few data are available for this children's intelligence test. Table 35 summarizes the results reported in two studies.

Correlations of IQ with teachers' ratings of their pupils' intelligence were .68 (full), .64 (verbal), and .53 (performance).

⁴⁰ J. I. Krugman, "Pupil Functioning on the Stanford-Binet and the Wechsler Intelligence Scale for Children," *Journal of Consulting Psychology*, Vol. 15, pp. 475-483, 1951.

Conclusions on Validity. On the whole, these results are encouraging; the correlation coefficients fall within the approximate range of indexes usually found for other widely used tests of intelligence, including the Stanford-Binet. These and similar data are, however, not yet definitive; the number of cases is small in both studies; and in one of the studies, grade level and age range were not sufficiently controlled. Further research is necessary at each of the age and grade levels for which the scale is intended before generalizations may be offered with assurance regarding the scale's predictive efficiency in respect to educational achievement.

Evaluation and Criticisms. On the whole, this intelligence scale for children is a useful addition to the very limited number of instruments now available for individual testing. It is to be expected that the deficiencies regarding standardization population and studies of reliability and validity will be remedied as the scale continues to be used for both practical and research purposes. Some psychologists, at this stage, question the wisdom of using this Wechsler scale as a substitute for the Stanford-Binet until more definitive data are obtained for the former.

Although one of the advantages originally claimed for this scale was that it did not use the mental-age concept, it has since been found desirable to supply mental-age equivalents.⁴¹ The mental age concept is an extremely useful one when interpreted by qualified psychologists. This concept should be made an integral part of the scale.

The relatively low reliabilities of the *subtests* indicate that there is no merit in merely deriving a test profile for purposes of diagnosis and guidance. The reliabilities of part-scores must be high before profiles can be used with confidence. The *total* verbal, performance, and full scores, however, have yielded reliability coefficients at a satisfactorily high level of confidence.

Considerably more research remains to be done on the predictive efficiency (validity) of this scale. Data available thus far show that IQ differences between it and the Stanford-Binet are significant enough to warrant caution, in spite of the generally high correlations found between them; and, it must be added, some of the coefficients

⁴¹ D. Wechsler, "Equivalent Test and Mental Ages for the WISC," *Journal of Consulting Psychology*, Vol. 15, pp. 381-384, 1951.

between the two scales are only moderate.⁴² The Stanford-Binet IQ's tend to be higher within the "normal" range, especially at the earlier age levels. Since the Stanford-Binet has been in use much longer, and has been found to have considerable value in schools and clinics, and has been widely used as a validating criterion, it is probable the obtained discrepancies between the two scales will stimulate research on and improvement of the more recent instrument.

The limits of the IQ values given by the Wechsler full scale are from 46 to 154. This means that the scale cannot be used with individuals who rank above or below these limits. In terms of total number of persons, the percentages of such cases will be very small; but in particular instances this can be a serious limitation.

⁴² J. J. Pastovic and G. M. Guthrie, "Some Evidence on the Validity of WISC," *Journal of Consulting Psychology*, Vol. 15, pp. 385-386, 1951. A particularly significant finding in this report is that, on the Wechsler scale, the mean performance IQ's are higher than the mean verbal IQ's. If these results are corroborated, they will raise some doubt over the applicability of the scale to the usual problems of predicting and evaluating educational achievement.

8.

INDIVIDUAL PERFORMANCE SCALES

ABOUT the time the final revision of Binet's own scale appeared, some psychologists in the United States had assembled a group of performance tests intended to meet practical problems in the study of human abilities and behavior. One of these was the Healy-Fernald group of tests, devised primarily to examine juvenile delinquents in an effort to determine their intellectual levels and personality traits revealed in the course of the examination.¹ Unlike tests which were subsequently developed and which are now in use, those of Healy and Fernald were not actually standardized in respect to administration and scoring. This group of tests provided the psychological examiner with situations wherein he could observe, evaluate, and interpret the testee's methods of solving problems and his behavior in test situations. The specific tests were selected on the basis of Healy's and Fernald's judgment and psychological insights as to what constitutes intelligent activity; beyond this the value of the results obtained with their test would depend upon the clinical acumen of examiners, since there were no norms based upon standardization procedures. While the Healy-Fernald tests are infrequently used today, they are important in the historical development of performance scales.

Several of the older and of the current scales will be described; these descriptions will be followed by a discussion of their uses and by a general evaluation of this type of instrument.

¹ W. Healy and G. M. Fernald, *Tests for Practical Mental Classification*, Psychological Monographs, Vol. 13, No. 2, 1911.

THE PINTNER-PATERSON SCALE OF PERFORMANCE TESTS

Contents. This group of performance tests, the first to be organized into a scale, is now of interest principally for its historical and background value.² It will also enable the student to see that many of the earliest types of performance tests have survived the years of experimentation and application and have been incorporated into scales now current, including the Bellevue.

Pintner and Paterson standardized some of the Healy-Fernald performance tests as well as several which had been devised by other psychologists and themselves. The final scale includes fifteen tests which can be presented without the use of language; nor do they require the use of language on the part of the subject. They are intended primarily for use with persons having serious hearing defects and for non-English-speaking individuals. These and similar performance tests have been found valuable as supplements to verbal tests of mental ability, and also with subjects who, though they are English-speaking, have speech defects or reading disabilities.

The subtests in the scale are described below.

(1) Mare and Foal Form Board. This, of the picture-puzzle type, is a pictureboard of a mare and foal, in color. Sections of the board are removed to begin with; the subject must replace them correctly. Score is based on time required and number of wrong moves.

(2) Seguin Form Board. This is a form board in which ten common geometric shapes are to be placed. Score is based on the shortest time required in three trials.

(3) Five-Figure Board. There are five geometric figures, each of which is divided into two or three parts. The pieces are to be fitted into their appropriate places. Score is based on time required and number of errors made.

(4) Two-Figure Board. There are two geometric figures, one cut into four sections and the other into five. These are to be correctly placed in two spaces. Score is based on time required and number of moves.

(5) Casuist Board. This form board—more difficult than the preceding ones—consists of four spaces in which twelve sections have to be fitted. Score is based on time required and number of errors made.

² R. Pintner and D. Paterson, *A Point Scale of Performance Tests*, New York: D. Appleton, 1917.

(6) Triangle Test. Four triangular pieces are to be fitted into the board. Score is based on time required and number of errors made.

(7) Diagonal Test. Five variously shaped sections have to be fitted into a rectangular form. Score is based on time required and number of errors made.

(8) Healy Puzzle A. This consists of five rectangular sections which are to be fitted into a rectangular frame. Score is based on time required and number of moves made.

(9) Manikin Test. Wooden legs, arms, head, and body are to be put together to make the form of a man. Score depends on quality of performance.

(10) Feature Profile Test. Wooden sections have to be put together to form the profile of a man's head. Score is based on time required.

(11) Ship Test (originated by H. A. Knox). This is a picture of a ship cut into ten sections, all of same size and shape, to be inserted properly in a rectangular frame. Score depends on quality of performance.

(12) Healy Picture-Completion Test I. This is a large picture from which ten small squares have been cut out. The missing parts are to be selected from among forty-eight squares identical in size. Score depends on the quality of completion within a limit of ten minutes.

(13) Substitution Test. A page of rows of geometric figures (five different shapes) which have to be marked with appropriate digits, to correspond with a key at top of page. Score is a combination of time and errors made.

(14) Adaptation Board. This is a form board having four circular blocks and holes; three are 6.8 cm. in diameter, while the fourth is 7 cm. The subject is shown that one block fits the larger hole. He is then required to keep his attention fixed and to fit this larger block into the correct space when the board is moved into four different positions. Score is based on the number of correct moves.

(15) Cube Test. Four cubes (one inch) are placed before the subject. With a fifth cube they are tapped in a specified order by the examiner. The subject is asked to imitate the order of tapping. The sequence becomes longer and more complex. Score is the number of sequences correctly imitated.

For general testing purposes, the authors of this performance scale recommend the use of a short scale which includes ten of the fifteen parts: namely, 1, 2, 3, 4, 5, 9, 10, 11, 12, and 15 of the foregoing list.

The age range of the Pintner-Paterson scale is from four years to fifteen. However, this does not mean that every test in the series has

discriminative value throughout this range. For example, the Seguin Form Board does not have value in general beyond age ten, while the Feature Profile test is not generally useful below age ten.

Scoring. Three different methods of scoring were provided by the authors: median mental age, point score, and percentile rank. For each test there is a separate table of mental age norms; the median of an individual's MA's on each of the several subtests is taken as the single MA to represent his general performance on the entire scale. In the point scale, the subject earns a total score in points for all the parts; the total score determines his MA, as indicated in a table of norms. By the percentile method, the subject's score on each of the several tests yields a percentile score; these can be combined to yield a single percentile rating for the whole scale. Of the three indexes, the median mental age has been used most widely with this scale.

Evaluation. During the many years that the Pintner-Paterson performance tests were used for clinical and experimental purposes, the following evaluations of them were widely accepted. They are more susceptible to practice effects, and chance successes are more frequent, than is the case with verbal tests; hence, the reliability coefficients of these performance tests are not as high as those of the verbal. This scale is useful primarily with young children, and with older children and adults who are mentally retarded or deficient. The scale has clinical significance, also, in the case of an older child when there is marked discrepancy in performance on the several subtests. The several parts of the performance scale examine processes that are more specific than those examined by verbal tests. This is indicated by the considerable scatter of ratings on the several parts and by the lower correlation coefficients found when ratings on each separate part were correlated with ratings for the whole scale, the range being from very negligible coefficients to fairly high, with a median at about .50.

Performance tests of the Pintner-Paterson kind correlate poorly or only very moderately with intelligence tests of the verbal kind as represented by the Stanford-Binet, when the group being studied is limited with respect to age or range of ability. For instance,³ when a group of gifted children was examined with both scales, the correla-

³ D. A. MacMurray, "A Comparison of Gifted Children and of Dull-Normal Children Measured by the Pintner-Paterson Scale as Against the Stanford-Binet Scale," *Journal of Psychology*, Vol. 4, 1937, pp. 273-280.

tion coefficient between the two sets of ratings was only .23. For a group of dull children who were likewise examined, the coefficient was .43. Assuming adequate reliability of both scales, these coefficients indicate that there is only little or moderate correspondence between them, regarding functions being tested. Furthermore, the very low coefficient of .23 suggests, also, that the performance scale

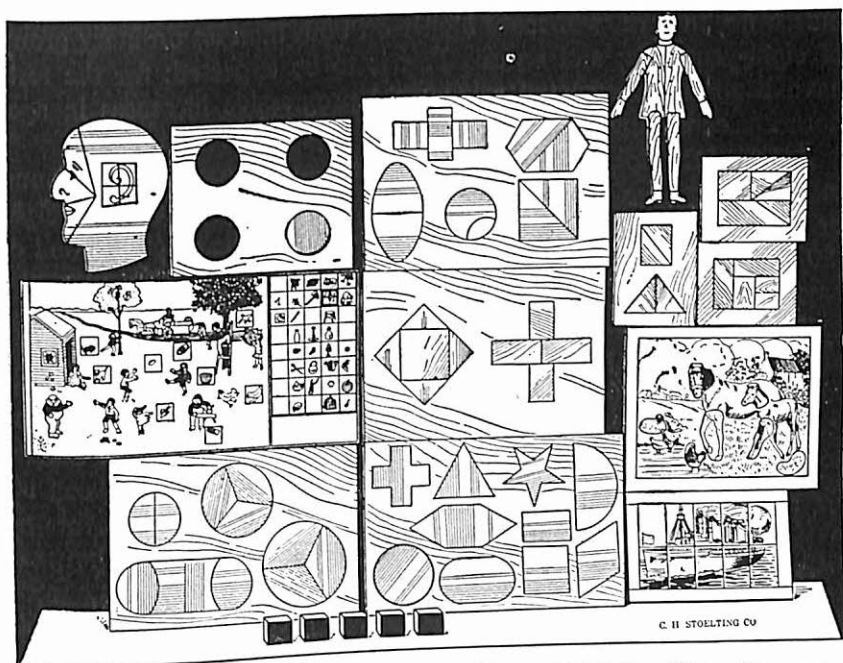


FIG. 8.1. *Pintner-Paterson Performance Tests*. C. H. Stoelting Company.
(By permission.)

is particularly inadequate for differentiating among performance levels of gifted children.

A fairly large number of studies have been published reporting higher correlations between performance-test ratings, obtained with the Pintner-Paterson and similar tests, and those obtained with revisions of the Binet scale. Coefficients of the order of .70 and .80 were not uncommon, whereas others were as low as .50. These coefficients, however, cannot be interpreted as necessarily indicating that there is a considerable community of function between these performance tests, on the one hand, and verbal tests of ability on the

other. Indeed, it appears that to a considerable degree the correlation coefficients are due to the wide age-range of the subjects tested, with the result that the coefficients reflect the fact that the psychological functions being tested by both types increase with age; that is, the results on both types of tests are to an appreciable extent the product of age. An ordinary group of ten-year-old children will get higher scores on both types of tests than will a similar group of nine-year-olds, who, in turn, will score higher on both than an ordinary group of eight-year-olds; and so on. This is to be expected; for the tests have been so constructed as to yield progressive increases in age-norms as chronological age increases.

Another example of the effect of age-range on correlations is found in the coefficients between intelligence ratings and height or weight or dentition. These, for a wide age-range, are in the neighborhood of .50 and .60, because in general, older children are taller, heavier, and have more permanent teeth. It would not be said, of course, that these coefficients indicate community of function between the psychological tests and the physical measures. But within a single age-group the correlation coefficients between these physical traits and intelligence test ratings drop down to negligible levels. Thus, when age is held constant, or very nearly so, the correlation coefficients between results obtained with the Pintner-Paterson and similar performance tests, on the one hand, and verbal tests, on the other, drop to between .40 and .60.

A factorial analysis of results obtained on thirty-four commonly used performance tests suggests one reason why there is only a low or moderate correlation between these and verbal tests of mental ability. This analysis appears to indicate that the principal factors measured by the performance tests may be identified as "spatial, perceptual speed, and induction."⁴ While the first two of these functions are involved to some extent in many verbal tests of ability, they are actually of relatively little significance there in the determination of an individual's rating.

These findings signify that verbal scales and those of the Pintner-

⁴ C. M. Morris, "A Critical Analysis of Certain Performance Tests," *Pedagogical Seminary and Journal of Genetic Psychology*, Vol. 54, 1939, pp. 85-105. "Perceptual speed" is the readiness to discover and identify perceptual detail (mainly visual). The "spatial" factor involves the ability to manipulate objects in space.

Paterson type may not be used interchangeably but should be used to supplement each other.

THE CORNELL-COXE PERFORMANCE ABILITY SCALE⁵

Contents. For this scale, the particular tests included were selected from a variety of sources. The authors carried out their own standardization and revised the directions for administering and scoring.

The tests included are the following:

(1) Manikin and Profile. These are already familiar to the reader. They are scored for accuracy and time required.

(2) Block Designs. These are the familiar Kohs colored-block designs, five of which were included. They are scored for accuracy and time required.

(3) Picture Arrangement. This includes ten series of pictures which, though different in subject matter, are the same in principle as those in the Bellevue scale. They are scored for accuracy only.

(4) Digit-Symbol. This type test is also familiar to the reader; for it is of the same kind as that included in the Bellevue and other scales. It is scored for accuracy and time required.

(5) Memory for Designs. This test includes five cards, on each of which is a geometric design. The subject is asked to reproduce each design after it has been shown for ten seconds. This type of test is similar to that used by Binet. The score depends upon quality of reproduction.

(6) Cube Construction. This test utilizes blocks some sides of which are painted while others are not. The examiner presents models of cube construction and asks the subject to duplicate them. The score depends upon both accuracy and time.

(7) Picture Completion. (This is an optional substitute for test 3.) The Healy Picture-Completion Test II was selected. The score depends upon accuracy only. (This test is the same in principle as Picture-Completion I; but its theme is different and on a higher level of difficulty.)

It will be noted that the Cornell-Coxe scale differs from other performance scales in that it does not include any form boards.

Scoring. In order that each test in the scale might contribute equally to the total score, the authors follow the common practice of converting raw scores into a type of standard unit.⁶ Total score is ob-

⁵ E. L. Cornell and W. W. Coxe, *A Performance Ability Scale*, Yonkers, N. Y.: World Book, 1934.

⁶ *Ibid.*, pp. 29 ff.

tained by adding the weighted scores for the several parts. A mental age is then obtained from a table of norms, extending from an MA of 4 years and 6 months to 16 years and 8 months.

Validity and Reliability. It appears from experimental data, cited by the authors, that they used the following criteria of validity: a "satisfactory" distribution of scores by school grades (increasing averages in successive grades); a distribution of scores that conforms well to the symmetrical bell-shaped curve; a high correlation between scores on each test in the scale with total scores. Although the correlation between total performance scores and chronological ages is .78, Cornell and Coxe do not regard CA as a validating criterion of first importance.

Reliability coefficients of the several parts of the scale varied from .66 to .89, while for total scores the reliability coefficient was .929 (125 cases). These data indicate a satisfactory degree of reliability for the scale as a whole.

The authors of this scale maintain that a performance scale should not be a substitute for those of the Binet type and other verbal tests, but should supplement them. They sought, therefore, to devise an instrument which should differ from these others in respect to functions tested. In this, they apparently succeeded fairly well; for although the correlation coefficient found between total performance score and Stanford-Binet (1916) mental ages was .79 for a wide age-range, when chronological age was held constant the partial correlation coefficient was reduced to .38. This means that in a group of about the same age the results of the two scales would indicate relatively low community of function.⁷

It appears also that the several tests within the scale have only very moderate community of function. Intercorrelations of the parts ranged from coefficients of .50 to .75, over the entire age-range; but when chronological age was held constant, the partial correlation coefficient varied from about .20 to .60.

In constructing their scale, Cornell and Coxe were interested primarily in developing a supplementary instrument. They say in this regard: "One important value of any scale supplementary to the

⁷ Practically the same results were obtained when the Cornell-Coxe scale was correlated with the National Intelligence Test, a group test of the verbal type. The simple correlation coefficient was found to be .74.

Binet scale lies in the fact that if the two scales used give different results, the psychologist's attention is directed toward discovering reasons for whatever differences may be found, and his analysis and interpretations are thereby enriched and tend to have greater validity."⁸ In view of the data on the degree of correspondence between results obtained with this performance scale, on the one hand, and those obtained with Binet revisions and verbal group tests, on the other, it is reasonable to conclude that the Cornell-Coxe scale serves the purpose for which it is intended.

THE ARTHUR POINT SCALE OF PERFORMANCE TESTS⁹

Contents. Form I of this scale is a restandardization of some of the tests used in the Pintner-Paterson, plus two other tests. The eight parts are: Knox Cube Test, Seguin Form Board, Two-Figure Form Board, Casuist Form Board, Manikin, Feature Profile, Mare

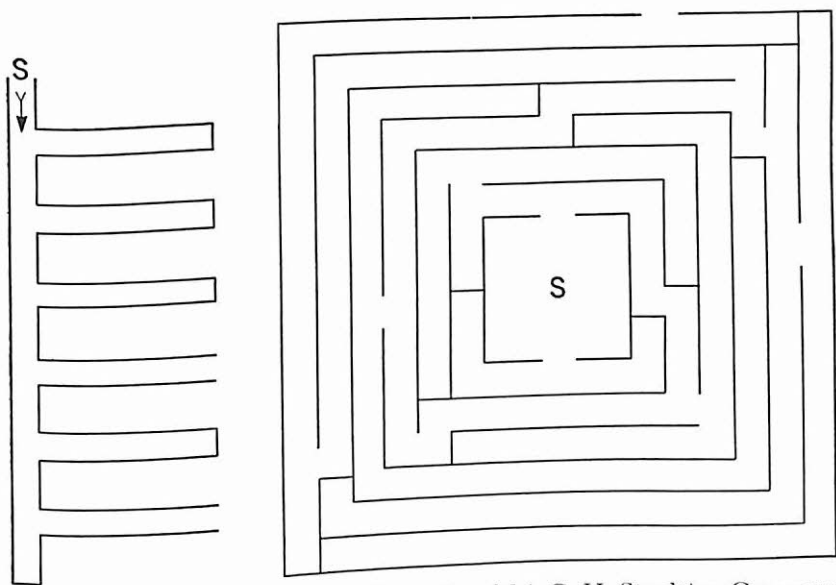


FIG. 8.2. Porteus Maze Tests—Years 5 and 14. C. H. Stoelting Company (By permission.)

⁸ *Ibid.*, p. 37.

⁹ G. Arthur, *A Point Scale of Performance Tests*, New York: The Commonwealth Fund, Vol. 1, 1930; Vol. 2, 1933; Vol. 1 appeared, also, in a revised edition in 1943.

and Foal, Healy Picture Completion I. The two additional are: Porteus Maze Test and Kohs Block Design Test.

The Porteus test consists of a series of mazes of increasing difficulty, each printed on a separate sheet. The subject is required to trace, with pencil, the course from entrance to exit. The Kohs test consists of the same set of blocks used in the Bellevue scale, but different designs are to be reproduced.

The purpose of the restandardization of the scale is to provide a more reliable and useful performance scale for clinicians. Form I is based upon results obtained with 1100 public-school children of middle-class families. The usual validating criteria were applied, such as parental occupation, age-grade distribution, significant increases in score in successive ages, and degree of correspondence with ratings obtained by means of other scales already considered to be acceptably valid (Stanford-Binet and Kuhlmann-Binet).

Scoring. An individual's score is variously determined on the several tests by number of successes, or time required, or degree of accuracy, or a combination of these. Each test yields a raw score which is converted into weighted score points.¹⁰ The raw score for each subtest is assigned a value proportional to the effectiveness of the subtest in differentiating between successive age-levels. The total of these weighted scores is converted into a mental age. The range of mental-age norms, in Form I, is from five and a half years to fifteen and a

¹⁰ The formula assumes that the value of a test, hence its weight in the total score, depends upon the extent to which it differentiates between successive age groups. The greater the difference between age-group averages of a test and the less the amount of overlapping of scores of two adjacent age-groups, the greater is the weight given to that test. The formula is:

$$DV = \frac{M_1 - M_2}{\frac{PE_1 + PE_2}{2}}$$

in which M_1 is the mean of the older of the two adjacent age groups; M_2 is the mean of the younger age group; the PE 's are the probable errors (middle 50 percent of the scores) of the two groups. DV stands for "discriminative value." Inspection of this formula shows that as the difference between means increases, the numerator grows larger and discriminative value increases. But the DV is also dependent upon the sizes of the probable errors, which indicate the amount of overlapping of scores between the age group. Thus, also, as the probable errors are smaller and, therefore, the overlapping less, the denominator is the smaller; and the fraction—hence the DV —is larger. See Arthur, *op. cit.*, Vol. 1, 1943, p. 39.

half. Arthur employs a statistical device to extend the norms downward by six months by using a constant monthly rate of decrease in score. This is a procedure of very doubtful validity because it assumes that rate of psychological development—as represented by these tests—is constant in the early years, whereas the prevailing conception among psychologists is that rate of development is most rapid in the earliest years and decreases as the child grows older.¹¹ The scale is similarly extended at the upper end by adding a constant value, in points, to provide hypothetical mental-age ratings beyond the norms derived by the actual standardization process.¹² When mental ages and intelligence quotients are obtained from norms thus extrapolated, the examiner must clearly realize that he is deriving indexes which do not necessarily have the same meaning as those found by actual standardization.

REVISED ARTHUR SCALE: FORM II

The purpose of the second form is to serve as an alternate when retesting, and for use with preschool children.

This version of the scale utilizes four of the test types already described, namely, Knox Cube, Seguin Form Board, Porteus Mazes, and Healy Pictorial Completion II. The only new type of material not thus far described is the Arthur Stencil Design. This test employs twenty designs, increasingly complex and more difficult to reproduce, which are presented singly. The testee is given six square colored cards and twelve colored stencils which are cut within square cards. Each design is to be reproduced by placing the appropriate cards and stencils, one upon another, so as to duplicate the original in both form and color. For example, a practice design requires merely that a red octagonal stencil be laid over a white card to get the desired result.

Reliability. Scores of Forms I and II were correlated at each age level from six to sixteen years. The coefficients, with CA constant, ranged from .55 ($PE \pm .06$) at age 8, to .70 ($PE \pm .05$) at ages 10

¹¹ Dr. Arthur herself recognizes this.

¹² The reader should contrast this with the method employed in constructing the Bellevue scale, remembering, however, that the Bellevue is designed primarily for adults while the Arthur scale is not. Compare also with method of deriving mental ages and intelligence quotients by means of the Stanford-Binet scale at levels of superior adults.

and 15. The median coefficient was .61 ($PE \pm .06$). As estimates of reliability, these coefficients are relatively low. The small number of cases in each age group, varying from 41 to 54, might account in part for these results.

That the scale is more reliable than the foregoing data suggest is indicated by the results of another study in which the subjects were 61 institutionalized mentally deficient boys whose mean IQ on the Stanford-Binet was 67. They were tested with the Arthur scale, then retested after an interval of two years, the correlation between the two sets of scores being .85. The coefficients for each of the parts varied from .69 to .80. The over-all results indicate satisfactory stability of relative rank, especially in view of the narrow range of ability within the group of subjects tested. It is to be noted, however, that the mean gain in the Arthur scale IQ was ten points, as contrasted with a mean loss of only one IQ point on the Stanford-Binet during the same interval.¹³

The mean gain of ten points may be attributed to one or both of two factors: (1) scores on the performance type of test are more susceptible to practice (learning) than are those on the verbal type; or (2) residence and training in a soundly conceived and operated institution encourages the development and utilization of the *potentialities* of the mentally deficient beyond levels attained under ordinary circumstances. The latter factor is not synonymous with specific practice and learning effect. It is, rather, a result of general training in more effective behavior; and, in some instances, also the removal of "blocks" that impair one's performance.

Validity. This performance scale was devised primarily as a clinical instrument to be used as a substitute for the Binet revisions in cases where a verbal type of scale is inappropriate, as in instances of language handicap, defects of vision or hearing, and inequality in development of an individual's verbal and nonverbal functions. Arthur, in her standardization procedures, has taken the position that the basic capacities demanded by the Binet tests and by her performance tests should be essentially the same; thus, the Kuhlmann and Stanford revisions of the Binet are two principal validating criteria

¹³ R. M. Patterson, "The Significance of Practice Effect upon Re-administration of the Grace Arthur Performance Scale to High Grade Mentally Deficient Children," *American Journal of Mental Deficiency*, Vol. 50, 1946, pp. 393-401.

used in constructing her scale. Accordingly, the main differences between the Arthur and the Binet revisions should be in the types of materials used to sample the psychological functions.

The extent to which the Arthur and Binet scales actually correspond may be inferred from a comparison of IQ ratings obtained by subjects examined with both, and from the correlation coefficients obtained between the two sets of ratings, taken separately for each age group.

In the first place, Arthur reports that the probable error (PE) of intelligence quotients was 4.97 points when the performance scale

TABLE 36
Correlations of Stanford-Binet IQ's
and Arthur IQ's

Age	N	R
5	35	.70 ± .06
6	54	.77 ± .04
7	50	.68 ± .05
8	44	.74 ± .05
9	41	.80 ± .04
10	40	.51 ± .08
11	44	.68 ± .05
12	31	.80 ± .04
13	27	.21 ± .12
14	27	.07 ± .13
15	16	-.10 ± .17

ratings were compared with those of the Kuhlmann-Binet, and 4.92 points when compared with the Stanford-Binet. What this means is that in fifty percent of the cases, the IQ differences were five points or less, while in the remaining fifty percent the differences were greater. The frequencies of the differences, however, decline markedly as the size of the differences increases beyond five points.

Correlation coefficients between Stanford-Binet (1916) IQ's and Arthur scale IQ's (Form I) are shown in Table 36.¹⁴ Most of these coefficients are rather high and noteworthy. Excepting at age ten, they show unusually high correspondence for these two types of scales between ages five and twelve. The correlations at the later ages, however, are so low and the probable errors (PE) so large that the coefficients

¹⁴ Calculated from data in Arthur, *op. cit.*, Vol. 2, pp. 54-61.

may be regarded as being zero for all practical purposes. If the table of coefficients is representative of the correspondence existing between Stanford-Binet IQ's and Arthur Performance IQ's at ages above twelve, then we must conclude that the latter scale has been inadequately standardized or is incapable of differentiating among individuals at the later age-levels in respect to the functions being measured.

Although the coefficients for ages five to twelve are quite marked, and in several instances high, they are, nevertheless, not close enough to unity (+1.00) to warrant the use of the Stanford-Binet and the Arthur scales interchangeably. For clinical purposes, the Arthur scale is valuable within the age range of five to twelve as a supplement to verbal scales of the Stanford-Binet type.

This conclusion is further supported by validating data obtained subsequent to the publication of Arthur's manual. These later studies, using both the 1916 and 1937 revisions of the Stanford-Binet, can be summarized as follows:

Stanford-Binet and Arthur IQ's correlate variously, from about .50 to about .80.

In a large majority of cases, the Arthur scale IQ's tend to be somewhat higher than the S-B at levels below 90 IQ.

At the levels above 90 IQ, the S-B tends to yield somewhat higher ratings.

The means of the differences between the IQ's of the two scales have been found to range from about 5 to 10 points.

While "discriminative value," regarded by Arthur as a most important criterion of validity, is no doubt significant, it does not in itself demonstrate that a scale is measuring the functions it has set out to measure.

There is, therefore, need for more definitive studies of the validity of the Arthur scale, using large enough numbers of subjects, including not only clinical and institutional groups, but a normal population sampling as well.

Arthur's own position appears to be that the appreciable extent of agreement between Stanford-Binet test results and those of her own performance scale indicates rather even development and manifestation of psychological functions in general. But, she believes, if

the results of these two scales disagree significantly in the case of a given individual, this is due to unevenness in development and expression of functions; or some complicating nonintellectual factors are responsible for the discrepancy.

In the case of any individual instance, the actual interpretation of performance test results, taken in conjunction with verbal test findings, will depend upon all information available with regard to the person concerned and upon the psychologist's interrelating of all relevant facts and data.

OTHER PERFORMANCE TESTS

It is not our purpose to present a description of all available tests of the performance type. We have presented several in some detail in order to acquaint the reader with their nature and their uses. Those scales described are typical. There are, however, several others which are intended to serve a special purpose; of these, three will be very briefly described.

The Ferguson Form Boards.¹⁵ The first description of these was published in 1920. They consist of a series of six form boards used as a unit and progressing in difficulty by fairly equal intervals. The tests were standardized upon 364 subjects ranging from children in grade one to college seniors.

Ferguson, apparently, used grade placement and school achievement as the principal evaluating criteria; for he reported correlations of his form board scores as follows: with grade placement, .81; with teachers' estimates of intelligence, .50; with class standing, .56.

Since their appearance in 1920, these form boards have been subjected to experimentation from time to time for purposes of revising procedure in administering and scoring, and providing more adequate norms. One of the most thorough revisions is that by Wood and Kumin (see footnote 15) who give norms for the ages of 7 years and 6 months to 17 years and 5 months. But, as is the case with many other tests of this type, a very large percentage of the standardization population consisted of individuals who had come to a guidance

¹⁵ G. O. Ferguson, "A Series of Formboards," *Journal of Experimental Psychology*, Vol. 2, 1920, pp. 47-58; L. Wood and E. Kumin, "A New Standardization of the Ferguson Formboards," *The Journal of Genetic Psychology*, Vol. 54, 1939, pp. 265-284.

clinic for assistance and who, therefore, may not be representative of the general population of their age groups. It is necessary to consider this fact when using and interpreting results of performance tests so standardized.

For present purposes, the fact of major interest is the extent to which these form boards and the scales of the Stanford-Binet type do or do not test common functions. Wood and Kumin report the following correlation coefficients: Ferguson score and Stanford-Binet (1916) mental age, simple correlations, .54 for boys and .55 for girls; partial correlations, when chronological age is held constant, .34 for boys and .47 for girls. These coefficients indicate a relatively low community of function between the two tests.¹⁶ In this respect, the Ferguson form boards are quite similar to nearly all the other performance tests presented in this chapter.

Kent-Shakow Form Board Series.¹⁷ This series of four form boards, devised by Kent and Shakow, was made available in 1925, as the Worcester Formboard Series. The series presents eight separate tasks graded in difficulty. In 1928 a modified series appeared, available in two forms: the Industrial Model and the Clinical Model which differ in respect to size, the former being the larger. The scale was first developed as a clinical instrument, with special reference to the needs of the Worcester (Massachusetts) State Hospital out-patient department. The standardization population consisted of 150 subjects from the age of six years upward, including adults. The early standardization, however, was very inadequate. Since these form boards were used frequently with adult subjects, better-standardized adult norms were published in 1939, based upon a population of average and superior adults numbering 355. The authors of this performance test do not attempt to validate it in terms of other established tests. Apparently, they are concerned mainly with providing a clinical device

¹⁶ The results were of the same general order of magnitude when the Kuhlmann-Anderson tests were correlated with Ferguson scores, namely, simple correlation coefficients of .47 for boys and .45 for girls.

¹⁷ D. Shakow and G. H. Kent, "The Worcester Formboard Series," *Pedagogical Seminary and Journal of Genetic Psychology*, Vol. 32, 1925, pp. 599-611; G. H. Kent and D. Shakow, "Graded Series of Formboards," *Personnel Journal*, Vol. 7, 1928, pp. 115-120; D. Shakow and B. Pazeian, "Adult Norms for the K-S Clinical Formboards," *Journal of Applied Psychology*, Vol. 23, 1939, pp. 495-502; W. R. Grove, "Modification of the Kent-Shakow Formboard Series," *Journal of Psychology*, Vol. 7, 1939, pp. 385-397.

which, they believe, measures manipulative skill and form analysis, and which, presumably, also provides a means of observing the subject's modes of approaching a problem.

In 1939 Grove published a modification of the Kent-Shakow series, Industrial Model, based upon an experimental group of 300 "native

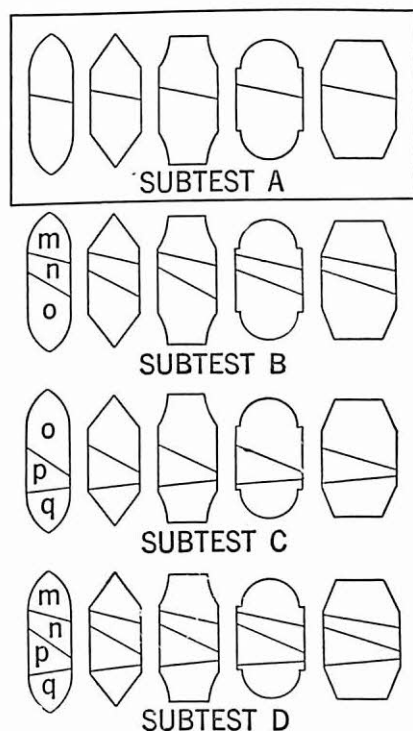


FIG. 8.3. Modified Kent-Shakow Form Board Series. (By permission of William R. Grove.)

born white adult male prisoners incarcerated in Western Penitentiary (Pennsylvania).” It is especially noteworthy that the scores obtained on this modified series of form boards, when correlated with Stanford-Binet (1916) ratings, yield a coefficient of only $.43 \pm .03$ (PE). The effect of age on this correlation is negligible, since it is based on a comparatively homogeneous group of about 300 adults. Thus, there is relatively little community of function measured by these two tests. Grove believes that his revised series measures ability to solve prob-

lems presented in the form of concrete spatial relations. It is reasonable to assume that the original Kent-Shakow tests measure the same functions as does the modified series, even though the technical terms used to name these functions are different.

The Carl Hollow Square Scale. This is a form-board test designed for use primarily with adults, though it is usable also with children over ten years of age. The test consists of a ". . . wooden panel in which is cut a $4\frac{1}{2}$ inch square hole, and 29 blocks of varying straight line geometric forms, each having both straight and beveled edges. There are long rectangles, short rectangles, three classes of right triangles, diagonally truncated long rectangles, diagonally truncated short rectangles, and overlapping rectangle-triangles."¹⁸ The problem for the testee is to fill the hole with sets of blocks, in a series of twenty tasks which become progressively more complex and difficult. Performance is scored on the basis of time required and number of moves made. Total scores may be converted into IQ's and MA's, or into percentile ranks.

This form-board test, its constructor believes, measures the following psychological processes: auditory memory (remembering principles and rules given in the instructions); visual memory (recall of partially repetitive patterns); observation and attention to detail; visual imagery involving synthesis and analysis (planning the placement of blocks without actual manipulation); learning (carry over from earlier tasks to subsequent ones).

It appears that the principal validating criterion employed was correspondence with results obtained with other tests having rather wide acceptance (Stanford-Binet, Kohs Block Design, Otis group test, Terman group test, Thorndike-McCall Reading Test). Correlation coefficients between these criteria and the form-board results varied, in the case of adults, from .50 to .80; while in the case of children over 10 years, the coefficients were from about .60 to .80. With adults, the factor of chronological age would not tend to produce spuriously high coefficients; but when the coefficients found with children are being interpreted, due regard should be given to the age range from 10 to 16 years (the theoretical beginning of adult level).

¹⁸ G. P. Carl, "A New Performance Test for Adults and Older Children: The Carl Hollow Square Scale," *Journal of Psychology*, Vol. 7, 1939, pp. 179-199.

A coefficient of reliability of .87 is reported for this test, indicating an acceptable degree of consistency that compares favorably with other tests of the verbal as well as the nonverbal type.

While the author grants that his performance test measures mental abilities which are involved more in the concrete and practical aspects of activity than in the abstract, he also maintains that it is more a measure of general than special ability. His view, presumably, is based on the correlations with the other tests of general ability already mentioned. This question will be dealt with in the section evaluating performance scales as a group.

FUNCTIONS TESTED BY PERFORMANCE SCALES

This discussion will supplement the analysis that was presented in connection with the nonverbal parts of the Stanford-Binet and the Wechsler scales. The reader should refer to those for more detail.

Since all performance tests involve visual perception and manipulation of objects, the number of types of items is relatively limited. It is not surprising, therefore, to find that the range of psychological functions is also restricted. This is one reason why the correlations between performance scales and scales of the Stanford-Binet type are not higher than they are, since the latter can sample a much wider range of functioning.

If the reader will re-examine the descriptions of the fifteen subtests in the Pintner-Paterson scale, and the few other types introduced in later scales, it will be readily apparent that they may all be classified in one of a few categories:

- geometric form boards, with variations, from the very simple to rather complex

- picture form boards (also known as picture completion) of various degrees of complexity

- block designs from simple to complex mazes of varying degrees of complexity

- recall of geometric designs

- picture arrangement

- block building

- cube sequences (imitating the order of tapping a series of cubes)

- digit-symbol

With the exception of the last two types, the performance items test visual perception plus more or less visual insight requiring analysis and synthesis. Performance on all the eight types also reflects motor speed in varying degrees; and performance on these is facilitated by visual imagery, that is, by the ability to analyze or synthesize a pattern imaginably before actually going through the movements.

Block building and cube tapping sequences are largely matters of imitation requiring the functioning of visual imagery and recall. The digit-symbol test, as already indicated in connection with the Bellevue, utilizes immediate rote recall and visual imagery.

In all of the performance tests, visual-motor integration, affecting the speed with which a person responds, is involved.

Also, many clinical psychologists hold that performance tests provide an estimate of the subject's attention span, especially in the case of mentally retarded and deficient individuals. Attention span, however, is not a process in the sense that visual analysis, memory span, etc., are. Attention is, rather, an *attribute* of the situation in which an individual is placed. If the testee is interested in the task at hand, and if the test is within the range of his apprehension, he will be attentive. If he does not understand the task, and if he is unable to make any progress with it and is confronted by repeated failure, he will very probably be inattentive.

It will be noted that these test items make few or no demands upon abstraction,¹⁹ concept formation, or the necessity of transcending the immediate concrete situation. For this reason, performance tests are regarded as having limited value as measures of general capacity, especially in the testing of individuals who are above average level.

EVALUATION OF PERFORMANCE TESTS

These tests were first constructed as *substitutes* for verbal scales of the Binet type. Many correlational studies showed, however, that it is sounder practice to regard the former as *supplements* to the latter. The reason for this interpretation is that, when allowance is made for the factor of chronological age, psychologists have found in almost every instance that the coefficients of correlation fall at .50 or lower—generally lower. Hence, although the two types of tests measure some functions in common, or are in other ways interrelated,

¹⁹ "Abstraction" means the separation of a quality, or an idea, or a principle from the organization and details of a concrete situation.

each type also measures functions different from those of the other type.

Performance tests have been found most useful with persons handicapped by language disabilities: the deaf, the non-English speaking groups, the illiterate, and those who have speech or reading difficulties.

These tests are valuable also in helping to identify children who are shy or inarticulate because of emotional reasons and who, therefore, may appear at a disadvantage on verbal tests of mental ability.

Performance tests, used together with the verbal type, are helpful in identifying the mentally deficient and mentally retarded with increased certainty. In cases involving diagnosis of mental deficiency, it is often desirable to supplement the Stanford-Binet, or a similar scale, with performance tests in order to check on the probable roles of the language factor, lack of cultural opportunities, and poor educational experience in order to estimate to what extent these might have adversely affected the testee's score on the verbal type of scale. If a significant difference is found between the two obtained ratings, further study of the individual is indicated before a conclusion is reached. It has often been found that mentally deficient and mentally retarded persons obtain somewhat higher ratings on performance than on the Stanford-Binet and similar scales. The differences, however, are not always significant enough to raise a question regarding the diagnosis. Furthermore, it is to be expected that, in many cases tested, the differences would be in the direction stated, since the performance tests were given for the very reason that the examining psychologist judged that the testee might be more successful with performance test problems.

Arthur, on the other hand, reported that for 435 clinic cases, having Stanford-Binet IQ's of less than 95, there was no group trend in the direction of higher IQ ratings on the Arthur Performance Scale, even among the duller individuals in the group.²⁰ She also found that for 60 mentally deficient cases (ages 15 to 20) who were routinely examined, the differences between verbal and performance IQ's were, on the whole, negligible. These are significant findings with regard to

²⁰ Grace Arthur, "An Attempt to Sort Children with Specific Reading Disability from Other Non-Readers," *Journal of Applied Psychology*, Vol. 11, 1927, pp. 251-264; "The Relative Difficulty of Various Tests for Sixty Feeble-minded Individuals," *Journal of Clinical Psychology*, Vol. 6, 1950, pp. 276-279.

group trends and group generalizations. But they do not alter the fact that in *some individual cases*, however few, a performance scale might yield sufficiently discrepant results to preclude a diagnosis without further study of the case. A discrepancy between an individual's rating on a performance scale and that on a verbal scale should be useful, rather than otherwise, in a clinical situation because then the psychologist must find the reasons for and seek an interpretation of the discrepancy. This necessity prevents the formulation of unwarranted conclusions, and it results in a fuller understanding of the person being examined.

Among the advantages reported for performance tests are these: (1) since they do not require the use of language, individuals do not "block" as a result of feelings of inadequacy due to lack of formal schooling; (2) since all elements of the problem are visually present, some individuals proceed with greater confidence.

Clinical psychologists are agreed that, where indicated, the use of performance scales can provide more information than just a rating in the form of a numerical index. These tests provide an opportunity to observe qualitative aspects of behavior under standardized conditions in a variety of problem-situations. A subject's approach to a problem might reveal, for example, a state of depression or agitation; hesitation or impetuosity; thoughtful deliberateness, bull-headed persistence, or easy discouragement; an insightful approach or one of haphazard trial-and-error.

Performance tests also have their disadvantages and limitations. As already stated, they are limited in range of mental functioning tested; hence they do not differentiate well among individuals of better than average levels. In fact, most performance scales thus far developed do not differentiate adequately among a large portion of the members of a representative population above twelve or thirteen years of age. On the whole, it is at the lower age levels and the lower mental levels that performance tests are most useful, in addition to their usefulness with persons having the handicaps already mentioned.

Geometric form boards, picture completion boards, object assemblies, etc., are within almost universal experience of American school children; and since they make demands upon the significant mental processes already indicated, they may be regarded as, in some degree, measures of intelligence at these earlier age levels. However, as performance tests go higher in age levels and difficulty levels, they present

problem-situations which are highly specialized and for the handling of which the subjects have by no means had a common background. This is especially true of the more complex and subtle form boards (e.g., Carl Hollow Square), performance on which is facilitated by training and experience in tasks requiring spatial perception, as in some types of engineering, cabinet making, and the like. Since these tests do not require, to a significant degree, the use of ability to make abstractions and to deal with concepts, they fail to measure some of the most important aspects of mental activity.

The reader will have noted that not all authors of performance tests agree as to what their characteristics should be. The Arthur scale was constructed to provide a nonverbal substitute for the Binet revisions. Hence it was expected to have a very significant correlation with these revisions, and standardization proceeded on that principle. The Cornell-Coxe scale is intended not as a substitute for Binet revisions but as a supplement to them. Thus, this performance scale was constructed on the principle that there should be a relatively low correlation between it and the verbal type of intelligence tests. Experimental evidence suggests that, at least in their present stage of development, performance scales are most properly and advantageously used in the manner advocated by Cornell and Coxe. For performance scales are instruments with which we may test development of insightful behavior involving visual perception rather than through the use of symbols (language and number) which are essential for abstractions, concept formation, ideational reasoning, and ability to deal with problems extending beyond one's immediate, concrete environment.

SCALES FOR INFANTS AND PRESCHOOL CHILDREN

IN THIS chapter we shall present several representative scales devised to evaluate mental development of individuals ranging in age from one *month* to six *years*. Some of these scales, for the greatest part, are not tests as that term is commonly understood. They are, rather, norms and inventories of development and behavior, grouped at their respective average age levels, derived from observation of children's behavior and from experimentation in a variety of situations. All are administered individually, of course.

GESELL DEVELOPMENTAL SCHEDULES

These are a product of systematic study of infants and young children at the Yale Clinic of Child Development. The first schedule,¹ an early effort published in 1925, provided rather crude norms at the following age levels: 4, 6, 9, 12, 18, 24, 36, 48, and 60 months.

At each level the inventory of activities was divided into four categories of behavior: (1) *motor*, (2) *adaptive*, (3) *language*, (4) *personal-social*. Although the normative schedules themselves have undergone considerable revision and refinement since their first appearance, these four categories, with some minor variations in terminology and analysis at times, have remained throughout. Motor behavior is said to be of value “. . . because it has so many neurological implications, and because motor capacities of the child con-

¹ A. Gesell, *The Mental Growth of the Preschool Child*, New York: Macmillan, 1925.

stitute the natural starting point for an estimate of his maturity." In adaptive behavior ". . . we reckon with the finer sensori-motor adjustments to objects and situations: the coordination of eyes and hands in reaching and manipulation; . . . the capacity to initiate new adjustments in the presence of simple problem situations which we set before the infant." Language behavior, broadly used, includes ". . . all visible and audible forms of communication, whether by facial expression, gesture, postural movements, vocalizations, words, phrases, or sentences. [It] includes mimicry and comprehension of the communications of others." Personal-social behavior ". . . comprises the child's personal reactions to the social culture in which he lives [bladder and bowel control, feeding abilities, sense of property, self-dependence in play, cooperativeness, responsiveness to training and social conventions]." ²

The Infant Schedule. The schedules of 1925 were followed by reports of further investigations upon which revised normative inventories of behavior were based. One schedule has been devised for the examination of infants between ages of four *weeks* and fifty-six *weeks*. At the four week level, the inventory of behavior includes analysis of head control, arm-hand posture, leg-foot posture, body posture and progression, regard, prehension, language and social behavior. At the fifty-six week level, the inventory includes the following categories: body posture and progression, prehension, manipulation and adaptation, language and social behavior.³ Each of these categories of behavior is evaluated, in the case of a particular infant, by observing him in a number of situations.⁴ Each situation is broken down into a number of possible activities detailing the manner in which the infant might respond. (See Table 37.⁵) Since the enumerated responses follow diverse trends with age, they have been designated as follows: (1) *decreasing trend*, if at ascending ages there is a progressive de-

² A. Gesell and C. S. Amatruda, *Developmental Diagnosis*, revised edition, New York: P. B. Hoeber, 1947, pp. 5-6.

³ A. Gesell and H. Thompson, *The Psychology of Early Growth*, New York: Macmillan, 1938, pp. 147 ff.

⁴ For example, activity with a ball, a bell, rattle, cubes, cup, spoon, form-board, mirror, boxes, pellet and bottle, dangling ring, paper and crayon; patterns of body posture; locomotion; spontaneous activities in various daily situations, such as in toilet, bath, and crib; kinds of play; responses to people; kinds of vocalization.

⁵ From Gesell and Thompson, *The Psychology of Early Growth*, New York: Macmillan, 1938, p. 127. (By permission.)

TABLE 37

Dangling Ring Behavior (4 weeks-28 weeks)

Situation: Dangling Ring (RD)

RD	Behavior Items	4	6	8	12	16	20	24	28
(1)	Regards after delay	77	54	64	65	27	13	14	5
(2)	Regards immediately	26	46	36	35	68	97	96	95
(3)	Regards momentarily	53	85	71	38	35	..	..	..
(4)	Regards prolongedly	47	43	29	62	87	47	38	5
(5)	Regards consistently	..	..	..	..	17	26	59	90
(6)	Disregards in midplane	77	39	46	46	14	..	..	..
(7)	Regards in midplane	29	61	54	54	86	..	..	..
(8)	Regards in midplane (long head)	22	25	12	50	83	..	..	..
(9)	Regards in midplane (round head)	32	75	70	56	88	..	..	..
(10)	Regards ring in hand	..	..	..	..	66	82	100	100
(11)	Regards string	..	..	..	..	7	13	46	53
(12)	Shifts regard	94	100	100	96	93	46	38	41
(13)	Shifts regard to surroundings . . .	75	68	61	35	13	16	14	5
(14)	Shifts regard to Examiner's hand	28	64	61	77	48	..	..	..
(15)	Shifts regard to Examiner	41	54	57	65	64	27	24	27
(16)	Shifts regard to hand	0	4	7	8	19	5	3	..
(17)	Follows past midplane	44	62	50	58	84	..	..	..
(18)	Follows past midplane (lg. h.) . .	20	33	25	37	83	..	..	..
(19)	Follows past midplane (rd. h.) . .	55	75	60	67	77	..	..	..
(20)	Follows approximately 180° . . .	16	43	46	50	68	..	..	..
(21)	Follows approximately 180° (lg. h.)	0	11	25	25	83	..	..	..
(22)	Follows approximately 180° (rd. h.)	36	55	55	61	62	..	..	..
(23)	Approaches	0	0	11	12	62	89	96	100
(24)	Approaches after delay	..	..	..	..	58	30	19	9
(25)	Approaches promptly	..	..	..	..	32	66	81	91
(26)	Arms increase activity	0	4	11	42	64	..	..	..
(27)	Arms separate	0	0	4	15	17	19	7	..
(28)	Approaches with one hand	0	0	4	12	20	24	39	55
(29)	Approaches with both hands . . .	0	0	0	0	50	76	82	77
(30)	Approaches with arms flexed . . .	0	0	0	12	44	60	54	14
(31)	Hands come together	0	0	0	8	20	38	11	5
(32)	Contacts ring	3	4	4	15	43	81	100	100
(33)	Dislodges ring on contact	3	4	4	8	20	35	28	5
(34)	Grasps	0	0	0	8	22	73	96	100
(35)	Grasps after delay if grasps . . .	..	..	..	..	..	75	46	14
(36)	Grasps interdigitally	..	..	..	..	..	61	45	7
(37)	Retains entire period	..	..	..	..	20	19	40	65
(38)	Holds with both hands	..	..	..	..	10	33	56	67
(39)	Hand opens and closes on ring . .	..	..	..	..	30	11	10	14
(40)	Brings ring to mouth	..	..	..	..	38	58	82	74

TABLE 37 (continued)

RD	Behavior Items	4	6	8	12	16	20	24	28
(41)	Free hand to midplane	..	..	..	..	25	51	56	84
(42)	Transfers	..	..	..	..	3	18	41	74
(43)	Drops	..	..	..	..	78	56	41	32
(44)	Drops immediately	..	..	..	..	42	32	7	0
(45)	Regards dropped ring if drops ..	..	..	..	..	10	37	43	100
(46)	(If drops) pursues dropped ring	..	..	..	..	7	16	29	100
(47)	(If drops) resecures dropped ring	..	..	..	..	7	5	29	60
(48)	Rolls to side	3	4	8	4	35	42	38	18
(49)	Frets	9	14	4	8	27	23	32	21

crease in percentage of infants showing that behavior; (2) *increasing trend*, if at ascending ages there is a progressive increase in percentage showing that behavior; (3) *focal trend*, if at consecutive ages there is an increase, followed by a decrease in percentage giving that response. The "increasing" and "decreasing" behavior items were allocated to age levels on the basis of fifty percent frequency. The "focal" behavior items were placed at age levels at which they are most frequently observed.⁶

Scoring. The infant's responses are scored plus or minus, depending upon whether or not he manifests the enumerated behaviors. The score on each item of behavior is then noted on a record sheet in accordance with the categories listed above. The infant's "distinctive" (modal) level of behavior is found by observation; from this level, responses showing greater or lesser degrees of maturity are counted; an algebraic sum of the deviating responses is found; this sum is then related to the "distinctive" level in order to determine whether the trend is in a plus or minus direction. Finally, a rating is assigned the infant in each of the categories, thus providing a profile of development. The following scores are illustrative of the ratings that might be found in a particular case.⁷

postural behavior	28— weeks
prehensory behavior	28+ weeks
perceptual behavior	28 weeks
adaptive behavior	28 weeks
language behavior	32 weeks

⁶ A fourth type of item was also found: those having a *fluctuating trend*; that is, having more than one focus. But they are disregarded in the scoring.

⁷ For a detailed account of scoring method, see Gesell and Thompson, *op. cit.*, pp. 209 ff.

Scoring of the items in the several categories as plus or minus demands considerable clinical experience in observing and evaluating infant behavior. Also, once having scored the items, an appreciable element of subjectivity enters into the reading and interpretation of the scoring record sheet.

Validity and Reliability. This developmental schedule was not subjected to the usual tests of validity because its builders had no criteria of maturity levels to use as a basis of comparison. They do, however, present a rationale of their schedule as the basis for its validity. "Fundamentally the validity of the schedule here offered depends on the validity of the norms, the legitimacy of the category classifications, the appropriateness of each item for the category to which it is allocated, the soundness of the concept of maturity level, and the justness of using a sample of the child's behavior to indicate that level. . . . Our conclusions regarding them [the foregoing issues] go beyond experimental data and are based on years of clinical experience. We are justified in claiming their general soundness and practical applicability until contrary evidence is revealed." ⁸

It is maintained, likewise, by the authors of the schedule that reliability of their schedule cannot be determined by the usual statistical methods presented in connection with other types of tests. Gesell and Thompson state, "It is the systematic error [of testing] which we have tried to reduce by basing the schedule on a carefully planned and controlled study of infant behavior . . . the accuracy of the schedule resolves in last analysis to the question of the accuracy of each item of the norms." ⁹ And they believe the percentages at different age levels passing and failing various behavior items, as derived from their normative studies, are reliable to a satisfactory degree.

The Preschool Schedule. Gesell and his colleagues have also developed a scale extending from the age of fifteen *months* to six *years*.¹⁰ Norms are provided for the following age levels: 15, 18, 21, 24, 30, 36, 42, 48, 54, 60, and 72 months. For illustrative purposes, the schedules at the two extremes are given below.

⁸ Gesell and Thompson, *op. cit.*, p. 218.

⁹ *Ibid.*, p. 219. This statement by Gesell and Thompson, as evidence of reliability, leaves much to be desired.

¹⁰ A. Gesell, et al., *The First Five Years of Life*, New York: Harper, 1940. (Reprinted by permission.)

15-Month Level

Motor:

Walks: few steps, starts and stops
Walks: falls by collapse
Walks: has discarded creeping
Stairs: creeps up full flight
Cubes: tower of 2
Pellet: placed in bottle
Book: helps turn pages

Adaptive:

Cubes: tower of 2
Cup and Cubes: 6 in and out cup
Drawing: incipient imitation stroke
Formboard: places round block
Formboard: adapts round block promptly

Language:

Vocabulary: 4-6 words or names
Jargon: uses
Book: pats picture
Picture card: points to dog or own shoe

Personal-Social:

Feeding: has discarded bottle
Feeding: inhibits grasp of dish on tray
Toilet: partial toilet regulation
Toilet: bowel control
Toilet: indicates wet pants
Communication: says "ta-ta" or equivalent
Communication: indicates wants (points or vocalizes)
Play: shows or offers toy to mother or examiner
Play: casts objects playfully or in refusal

72-Month Level

Motor:

Jumps from height of 12", landing on toes only
Advanced throwing
Stands on each foot alternately, eyes closed
Walks length of 4 cm. board
Copies diamond

Adaptive:

Builds 3 steps with cubes
Draws man with neck, hands on arms, and clothes

Draws man with 2-dimensional legs
Copies diamond
Adds 9 parts to incomplete man
Discriminates 5 weights, no error
Detects missing parts of pictures
Repeats four digits
Gives correct number of fingers on single hand
and on both
Adds and subtracts within five

Language:

Binet items used here

Personal-Social:

Ties shoe-laces
Differentiates A.M. and P.M.
Knows right and left or complete reversal
Recites numbers up to the thirties

Scoring. These schedules are not scored quantitatively. They are a general clinical guide intended for use in estimating the developmental status of a given child in respect to the four designated categories of behavior.

Validity and Reliability. In respect to validity and reliability, Gesell and his collaborators take the same position, presumably, as that quoted in connection with their other schedule; for no statistical evidence is provided beyond percentages passing at the several age levels. "Application of the schedules is a simple matter of determining how well a child's behavior fits one age level constellation rather than another, by the method of direct comparison. . . . There is nothing mathematical in this determination, neither is there anything mystical about it. It amounts to matching, which is neither calculation nor intuition; . . ." ¹¹ Performance and developmental status are reported separately for each of the four categories of behavior in terms of the four approximate age levels.

Evaluation. Inspection of these two developmental schedules reveals that each is a combination of some aspects of mental development

¹¹ *Ibid.*, p. 320. While we believe there should be nothing mystical about validity and reliability of a psychological scale, we also believe that subjective comparison and intuition should not be substituted for scientific determination of reliability and validity.

(as usually understood), motor development, sensory development and perception, and development of personal habits (often called social development).

The schedule for infants (4 weeks to 56 weeks of age) has value for the experienced psychologist because it provides means, experimentally and clinically derived, of estimating, nonquantitatively, specified aspects of a child's development within the first year of life. But as psychological tests, this schedule does not satisfy the demands of standardization in terms of norms, reliability, and validity. The population sample was small (49 boys and 58 girls) and restricted (from a homogeneous middle-class background). It yet remains for some psychologists to subject the schedule to rigorous studies of reliability and validity before it can be used with considerable confidence for predictive purposes. On the positive side, it can be said that superior experimental techniques were used, and a great deal of careful observation, experience, and behavioral insight went into the derivation of the developmental schedule. For these reasons, when applied by skilled observers, it is useful in appraising an infant's developmental status as it appears *at the time of the examination*.

For the reasons stated above, the second schedule of development and behavior (for ages 15 to 72 months) is of questionable value. For this group of children, especially those two years of age and older, there are other scales that have been standardized and that can be used with more confidence. With some of these, the reader is already familiar (e.g., the Stanford-Binet); others will be described in the following pages.

MINNESOTA PRESCHOOL SCALE¹²

This scale, in two forms, is an adaptation and restandardization of test items chosen from the earlier work of a number of psychologists, plus some original additions. It is designed for use with children from age eighteen months to six years.

The scale includes the following twenty-six tests: pointing out parts of the body; pointing out parts in pictures; naming familiar objects; copying a circle, triangle, and diamond; imitative drawing (vertical

¹² Developed by F. L. Goodenough, J. C. Foster, and M. J. Van Wagenen. Published by Educational Test Bureau, 1932 and 1940. The revised test manual, 1940, is by F. L. Goodenough, K. M. Maurer, and M. J. Van Wagenen.

and horizontal strokes and a vertical cross); block building; response to pictures; Knox cube imitation (tapping a series of cubes in a given order); obeying simple commands; comprehension ("What should you do when you are hungry?"); discrimination of geometric forms; naming objects from memory, recognition of forms, color naming; tracing a form; picture puzzles (object assembly); incomplete pictures; digit span; picture puzzles, diagonal series (more difficult object assembly); paper folding; absurdities (verbal); mutilated pictures; vocabulary; word opposites; imitating position of clock hands; speech (length of sentence spoken by child during examination).

Scoring. The raw score is converted into what is known as a C-score, which in turn can be converted into an IQ equivalent by means of tables provided in the manual.¹³ The Minnesota scale also provides for another type of score known as "percent placement," which is defined as the percentage of the difference between the score of the most backward and the score of the most advanced child likely to be found in a representative group of a thousand children of similar age. Thus, if the lowest C-score in a given group is 50 and the highest is 110, the range is 60 points. A child who gets a C-score of 65 is 15 units, or 25 percent ($15/60$), above the lowest score in the group. His "percent placement" score, then, is 25.

The norms of this scale are so arranged that it is possible to obtain three separate scores for children above thirty months of age: a verbal, a nonverbal, and a total score. For a child under thirty months of age only the total score is used because the authors of the scale were unable to work out a system of differentiated scoring for these earlier levels. A rough analysis is possible, however, to determine whether a pronounced difference between verbal and nonverbal responses exists. If such a difference is found at any age within the range of the scale, then, as the case may be, handicap or acceleration in respect to language or perceptual-motor ability may be inferred.

¹³ The C-score ". . . represents the difficulty of the tasks with which [a child] may be expected to succeed in 50 percent of his trials" (*Manual*, p. 91). It is a form of "absolute scaling," the units of which, presumably, increase by steps that are approximately equal in difficulty. It is a variation on the familiar standard-score technique. For a more detailed description of the C-scores and the basis of determining IQ equivalents, see F. L. Goodenough and K. M. Maurer, *The Mental Growth of Children From Two to Fourteen Years*, Minneapolis: University of Minnesota Press, 1942, Chapter IV.

Validity and Reliability. The Manual of the Minnesota scale does not provide data specifically designated as evidence of validity. We may infer, however, that the authors regarded the following facts as their basis of validity: (1) the adaptation and use of types of test items considered by many psychologists, over a period of years, to have validity; (2) a standardization group of 900 children, ranging in age from eighteen months to six years (100 in each of nine half-year age groups), who were balanced equally as to sex and whose

TABLE 38

*Correlations Between Minnesota Preschool and
Stanford-Binet IQ's*

Age in Months at Taking Minnesota *	Correlations	
	1916 S.-B.	1937 S.-B.
Under 36	.45	.21
36-47	.64	.61
48 and over	.65	.68

* The number of cases in each group was large, ranging from 141 to 841.) From Goodenough and Maurer, *op. cit.*, Part II.

fathers were representative of the distribution of occupational levels in the general population.

In another and more recent publication, data relevant to the scale's validity are available.¹⁴ Children who had been tested originally with the Minnesota scale at various ages during their preschool years were retested with the 1916 Stanford-Binet, in some instances, or with the 1937 revision in others. The intervals between tests and retests varied from a few months to about ten years.

When the 1916 revision was used in retesting and the results were correlated with the total scores of the Minnesota, the range of coefficients for the various groups was from a low of .25 to a high of .75.

When the 1937 Stanford-Binet was used in retesting, the correlations with original Minnesota scale total scores yielded coefficients from .15 to .76.

Table 38 shows the median correlations between original Minnesota IQ equivalents, found at various ages, and the retest Stanford-Binet IQ's at ages ranging from 4½ to 13½ years.

¹⁴ Goodenough and Maurer, *op. cit.*, Part II.

If we accept the Stanford-Binet scales as significant criteria of validity, as many psychologists have in actual practice, then we must conclude that the Minnesota scale has low validity below the age of thirty-six months, but that it has much greater validity for individuals above three years of age.

In this connection two considerations must be kept in mind. First, it has been found that all scales devised for use with children below the age of eighteen months show a low or very moderate correlation with retest results in later childhood. The probable reasons for this fact will be presented later. Second, the correlation coefficients between results of testing and retesting the same subjects with the *same* scale tend to decrease somewhat as the time interval between examinations increases.

Reliability data of the Minnesota scale are variable at the different age levels. The coefficients between the C-scores on the two forms of the scale (that is, the test-retest method), with intervals of one to seven days, were: .68 to .94 for the verbal tests, .67 to .92 for the nonverbal tests, and .80 to .94 for the combined total scores. The average reliability coefficients for a single form, within an age-range of six months, were: .86 for the verbal, .82 for the nonverbal, and .89 for the total scores.

CATTELL DEVELOPMENTAL AND INTELLIGENCE SCALE¹⁵

This scale, of superior merit, covers the range from two to thirty months. Its test items are adaptations of many which were developed and included in earlier tests, notably those of Gesell and his associates. Cattell states the scale “. . . has been so constructed as to constitute an extension downward of Form L of the Stanford-Binet tests. Between the ages of twenty-two and thirty months Stanford-Binet items are intermingled with other items. Thus, using the infant test items for the early months and the Stanford-Binet tests for the older ages with a mixture of the two between, one continuous scale from early infancy to maturity has been attained.”¹⁶ The test items are grouped at age levels as they are in the Stanford-Binet. Groupings are provided at each month from two through twelve; at two-month intervals in the second year; and at twenty-seven and thirty months.

¹⁵ Psyche Cattell, *The Measurement of Intelligence of Infants and Young Children*, New York: Psychological Corporation, 1940.

¹⁶ *Ibid.*, p. 24.

The following three age levels illustrate the nature of the items and their arrangement.

Two months

- (1) Attends voice
- (2) Inspects environment
- (3) Follows ring in horizontal motion
- (4) Follows moving person
- (5) Babbles
- (Alt. a) Follows ring in vertical motion
- (Alt. b) Lifts head in prone position

Ten months

- (1) Uncovers toy
- (2) Combines cup and cube
- (3) Attempts to take third cube
- (4) Hits cup with spoon
- (5) Pokes fingers in holes of peg board
- (Alt.) Picks up spoon before cup

Thirty months

- (1) Differentiates bridge from tower
- (2) Imitates drawing lines and circles
- (3) Stanford-Binet three-hole form board rotated
- (4) Folds paper
- (5) Stanford-Binet identifying objects by use
- (Alt. a) Identifies pictures from name
- (Alt. b) Concept of one

The scale was standardized by longitudinal testing; 1346 examinations were made on 274 children at the ages of three, six, nine, twelve, eighteen, twenty-four, thirty, and thirty-six months.¹⁷

In the process of standardization, it was Cattell's purpose, among other things, to improve on earlier scales by: (1) improving objective procedures for administering and scoring; (2) eliminating items of the "personal-social" category, which are markedly influenced by home training; (3) eliminating items which are indicators of large motor control; (4) providing more accurate age-scaling; (5) provid-

¹⁷ It should be noted that while the scale was standardized only on these age levels, groups of items are provided at certain age levels between them. The placement of items between the standardization levels was estimated. Cattell states, however, that the indications are the scale may be used with only a little less accuracy with children between the standardization ages. At the same time, she urges the exercise of caution in interpreting test results at the ages between the standardization levels.

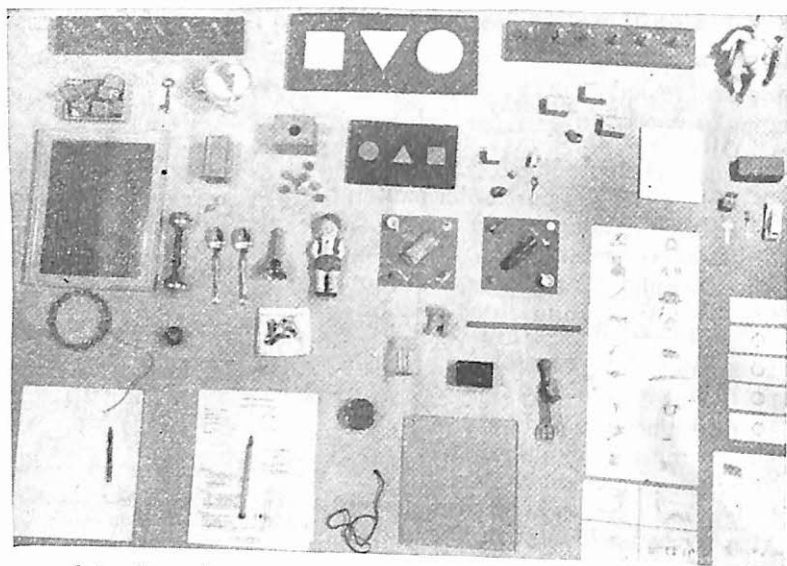


FIG. 9.1. Complete set of material for administering the infant tests. From P. Cattell, *The Measurement of Intelligence of Infants and Young Children*. Psychological Corporation. (By permission.)

ing an adequate age-range so that continuity of development can be studied; (6) providing a more nearly equal distribution of items over the age range covered.



FIG. 9.2. *Regards Cube*. Age 3 Months.

Material: A one-inch cube painted bright red.

Procedure: As the child is sitting in an upright position before the table, the cube is placed on the table within easy view of him. The cube may be tapped on the table or moved about to attract the child's attention.

Scoring: Credit is given if the child observes the cube. His eyes must remain on or return to the cube after the examiner has removed his hand. In other words, the examiner must make sure that it is the cube and not his hand which is observed.

From P. Cattell, *The Measurement of Intelligence of Infants and Young Children*, Psychological Corporation. (By permission.)

FIG. 9.3. Picks Up Spoon. Age 5 Months.

Material: Teaspoon.

Procedure: The spoon is placed directly in front of the child (sitting position) within easy reach.

Scoring: Credit is given if the child makes a definite effort to reach for and pick up the spoon and succeeds, but if the spoon is picked up by reflex closure of the hand on chance contact, it is not credited. Accurate reaching, however, is not to be expected at this age.

From P. Cattell, *The Measurement of Intelligence of Infants and Young Children*, Psychological Corporation. (By permission.)



Scoring. The method of scoring is the same as that used with the Stanford-Binet scale. Each item is scored as either plus or minus; no partial credits are given. Since there are five items at each age level, the credit given for each item passed is one-fifth of the interval cov-



FIG. 9.4. Places Round Block in Form Board. Age 16 Months.

Material: The form board is similar to Gesell's. It is made of a three-eighths-inch board 36×16 cm., stained dark green. Three holes are cut in the board equidistant from each other and from the edges. From left to right the holes are a circle 8.7 cm. in diameter; an equilateral triangle,

with sides 9.3 cm., and a square with sides 7.5 cm. The inserts are made of wood 2 cm. thick and painted white. The circle is 8.5 cm. in diameter, the sides of the triangle 9 cm., and those of the square 7.3 cm.

Procedure: The form board is placed before the child with the circle on his left and the base of the triangle toward him. The circle is placed in its recess and the child is allowed to take it out, then he is asked (with appropriate gestures) to "Now put it back."

Scoring: Credit is given if the child replaces the round block. If it is done with an evidently purposeful act, one trial is enough, but if there is some doubt as to whether or not it was a chance replacement, no credit should be given unless it is placed a second time. (Credit is given for replacing the round block in the reversed board at eighteen months.)

From P. Cattell, *The Measurement of Intelligence of Infants and Young Children*, Psychological Corporation. (By permission.)

ered by the particular series of tests. Thus, in a series of items spanning a one month interval, each item passed carries credit of .2 of a month; when the interval is two months, the credit per item is .4 of a month; with a three month interval, it is .6 of a month. The Cattell scale, like the Stanford-Binet, uses a basal age and sums up the credits at higher levels to obtain a mental age, and from that an IQ.

Validity and Reliability. Although percent passing each item at successive ages was used as evidence of validity, the principal criterion was the correlation between Cattell scale IQ ratings, obtained up to the age of thirty months, and Stanford-Binet (1937) IQ's obtained

TABLE 39
Validity Coefficients: Cattell and Stanford-Binet Scales¹⁸

No.	Ages at Examinations				Coefficients
42	3 mos.	and	36 mos.		.10 ± .10
49	6	"	"	"	.34 ± .08
44	9	"	"	"	.18 ± .10
57	12	"	"	"	.56 ± .06
52	18	"	"	"	.67 ± .05
52	24	"	"	"	.71 ± .05
42	30	"	"	"	.83 ± .03

with the same children at the age of thirty-six months. Table 39 shows the validity coefficients.

It is obvious that, accepting the Stanford-Binet as the criterion, the coefficients are practically negligible for tests given during the first nine months of life. In this respect they are much the same as other current scales. For the later ages, up to thirty months, the coefficients increase very appreciably and are on the whole superior to those found with most other scales designed for these age levels.

In spite of the low predictive value of the coefficients at the earlier ages, Cattell has found, from study of individual cases, that the tests may be of considerable assistance to the clinician in appraising infants who are marked deviants from the norm. This is the case especially with infants who get a high quotient; for, Cattell reports, they have

¹⁸ From Psyche Cattell, *The Measurement of Intelligence of Infants and Young Children*, New York: Psychological Corporation, 1940, p. 49. (By permission.)

appreciably better than average chances of earning a high rating at the age of two or three years.

Reliability of the scale was calculated by the odd-even procedure and corrected by the Spearman-Brown formula. Coefficients ranged from a low of $.56 \pm .05$ at the age of three months to a high of $.90 \pm .01$ at eighteen months. The median coefficient was $.86 \pm .02$. These coefficients compare favorably with those found for other scales.

MERRILL-PALMER SCALE OF MENTAL TESTS¹⁹

Although the norms of this scale are based upon 631 cases ranging in age from eighteen to seventy-seven months, its author does not recommend its use with children below twenty-four months or above sixty-three months of age.

The scale consists of ninety-three items arranged in order of difficulty. There is no attempt to group the tests according to types of function or behavior involved. The age norm (called "age at par") for each item is given, this being the age at which fifty percent of the children were successful. Although there are ninety-three items, there are only thirty-eight *different* items. Some (twenty-one) recur several times, at different age levels; at later ages a higher level of performance is required (in terms of quality or quantity of response, or in rate of activity) if credit is to be earned. These are called "variable-score" tests by the author. Other items (seventeen) occur only once in the scale; they are called "all-or-none" tests.

The scale tests some language (e.g., "What runs?" "What scratches?") These are known as action-agent tests. Also simple questions, like "What does a doggie say?"; manipulation of the body (e.g., opposition of thumb and fingers, crossing feet); motor skills and coordination (e.g., throwing a ball, buttoning); visual insights (e.g., building with blocks, copying a circle and a cross, completing form boards and picture puzzles); and recognizing familiar objects and colors.

Scoring. A point or raw score is obtained first. This may be converted into one or more of several relative indexes: namely, (1) men-

¹⁹ Rachel Stutsman, *Mental Measurement of Preschool Children*, Yonkers, N. Y.: World Book, 1931.

tal age, (2) standard deviation on the basis of point scores, (3) standard deviation on the basis of mental ages, (4) standard deviation on the basis of IQ's, (5) and percentile ranks on the basis of raw scores. It is suggested, however, that use of the IQ itself is inadvisable in connection with the Merrill-Palmer scale because its IQ deviations are not the same or close enough in size at the various age levels.

Stutsman provides and suggests the use of a "Guide for Personality Observations" in connection with this scale. While these observations do not affect the scoring, they are, nevertheless, useful to the clinician in interpreting a child's responses during the examination. The following traits are observed and rated, as they are manifested during testing: self-reliance, self-criticism, irritability toward failure, degree of praise needed for effective work, initiative and independence of action, self-consciousness, spontaneity and repression, imaginative tendencies, reaction type (slow and deliberate, calm and alert, quick and impetuous), speech development, dependence on parent, and other observations. Value of these observations and ratings, obviously, will be dependent on the skill and experience of the examiner. These and similar observations, as already pointed out, are desirable, in fact essential, in the complete report on and evaluation of any individual's test performance.

Validity and Reliability. Criteria of validity were those generally used: (1) known groups, (2) ratings by nursery school staff, (3) small overlapping of distribution of total scores between age groups, (4) correlation with chronological age ($r = .92 \pm .004$), and (5) correlation with the Stanford-Binet ($r = .79 \pm .019$; 159 children in the standardization group, between 3 and 6 years of age). The correlation coefficient for the last criterion must be interpreted in the light of the fact that the age range was three years.

In the guide describing the Merrill-Palmer scale and its standardization, no data on its reliability are provided. Subsequent studies, however, furnish information from which we may infer its reliability. When Stutsman retested a group of 77 children (ages 2 to 5 years) after an interval of two months, she found a correlation coefficient of .72 between the two sets of scores. Wellman, retesting a group of 44 children (ages 20-62 months) after an interval of one week,

found a correlation coefficient of .92 between the scores of the first and second tests.²⁰

EVALUATION OF SCALES FOR INFANTS AND PRESCHOOL CHILDREN

Technical Problems. Time or speed scores should not be used in tests for these age levels. The measurement of rate of performance is inadvisable and can be misleading, for at least two reasons: (1) speed of performance has not yet become a motivating factor in very young children; (2) the shifting attention of children at these age levels can obscure their true levels of skill and insight.

The grouping of test items according to types of activity, as in the Gesell schedule, has the advantage of readily indicating functions that are retarded and those that are accelerated in the case of the child being examined. While this kind of analysis is not so immediately apparent in an age scale, like Cattell's, it is nevertheless possible.

Since most of the usual validity criteria are not available in standardizing tests for infants and young children, this technical problem can be solved only through longitudinal studies, following up the same individuals over a considerable span of years, and correlating early test performance with later acceptable criteria of validity. Some efforts have been made in this direction.²¹

Determination of item and subtest reliability, within a scale, presents difficulties too. If the odd-even method is used, the results can be affected by the fluctuating attention of the subject. If the test-retest method is used, the results can be affected by the irregularity of

²⁰ B. L. Wellman, *The Intelligence of Preschool Children as Measured by the Merrill-Palmer Scale of Performance Tests*, University of Iowa Studies in Child Welfare, Vol. 15, No. 3, 1938.

²¹ See V. L. Nelson and T. W. Richards, "Studies in Mental Development: I. Performance on Gesell Items at Six Months and Its Predictive Value for Performance on Mental Tests at Two and Three Years," *Journal of Genetic Psychology*, Vol. 52, 1938, pp. 303-325; II. "Analysis of Abilities Tested at the Age of Six Months by the Gesell Schedule," *ibid.*, pp. 327-331; III. "Performance of Twelve-Months-Old Children on the Gesell Schedule and Its Predictive Value for Mental Status at Two and Three Years," *ibid.*, Vol. 54, 1939, pp. 181-191; "Abilities of Infants During the First Eighteen Months," *ibid.*, Vol. 55, 1939, pp. 299-318. Also, N. Bayley, "Consistency and Variability in the Growth of Intelligence from Birth to Eighteen Years," *ibid.*, Vol. 75, 1949, pp. 165-196.

growth tempos, when the time interval is significant. Significance of the time interval varies with the age of the subject: the younger the child, the shorter the significant interval. The desirable procedure would be to make retests within a week.

Although most scales for these early age levels extend upward beyond the two and three year levels, many psychologists recommend the use of the Stanford-Binet from age two, because of its more adequate standardization. Since the Cattell scale is an extension downward of the Stanford-Binet, and since it overlaps with the latter, it is a sound alternate for the Stanford-Binet. The Merrill-Palmer has also been found to be quite useful to the age of three or three and one-half years.

Uses. Psychological tests for children at these ages are used for two main purposes: (1) to determine a child's developmental status, with respect to the functions being evaluated, at the time of examination; and (2) to predict, so far as possible, future developmental and intellectual status. Most psychologists agree that the first purpose is reasonably well satisfied. With regard to the second of these purposes, the infant scales, with one possible exception, used with subjects below the age of fifteen or eighteen months have not proved adequate; for when test ratings obtained in about the first year and a half of life have been correlated with subsequent ratings obtained with the Stanford-Binet and other scales, the correlations found were so low as to be negligible. In fact, even an occasional small negative coefficient has been found. Therefore, when a child is examined within the first eighteen months of life, for purposes of predicting his future mental development and status, little weight can be placed upon the numerical rating, except in the cases of infants who deviate markedly from the average, in either direction.

Cattell's scale is superior to most others in this respect. The first three coefficients (.10 to .34) in the table showing validity coefficients of this scale are characteristic of those generally found for the first year of life. But the remaining coefficients of validity are higher than those found for other preschool tests. In general, predictive value of scales for preschool children increases after the age of eighteen months or two years. While the many correlational studies published on preschool groups aged eighteen months or more show coefficients varying over a considerable range, many of them fall in the .40's,

.50's, and .60's, with relatively few others higher or lower.²² In general, the higher the preschool age at the initial test, the higher will be the relationship of initial score to scores on retests.

In one instance, at least, psychologists have been concentrating upon devising scales to test very young infants exclusively—between the ages of 4 and 36 weeks—when adoptions are made most frequently. These scales, known as The Northwestern Infant Intelligence Tests, emphasize the child's adaptation to the physical and social environment. Their ultimate value, too, will depend upon their demonstrated predictive efficiency.²³

In spite of their low predictive value for very young infants, available scales are of assistance to an experienced clinical psychologist in appraising a child's behavioral and mental development when attention is given to analysis of performance on the various parts rather than to numerical scores alone, and when the analysis is used in conjunction with other clinical data. Developmental and intelligence tests for preschool children must be used with more than ordinary precaution; for the value of the findings is exceptionally dependent upon the skill of the examiner in eliciting the child's best efforts and in being able to appraise his general behavior during the examination session.

There are several reasons why results of tests in infancy and the earlier preschool periods do not have more value in predicting future mental status. First, resistance to examiner, shyness, failure to exercise maximum effort, and other emotional conditions are undoubtedly operative in some instances. More important and fundamental, however, is the fact that there are changes and irregularities in the tempo of development of numerous young children. It has been found that successive examinations of individual infants show fluctuations between two or three levels; or they may show a consistent trend downward or upward before leveling off to variations within a relatively narrow range of ratings.²⁴ These fluctuations and trends in rate of development may be due to changes in mental organization; that is,

²² For a compact summary of these studies, see K. M. Maurer, *Intellectual Status at Maturity as a Criterion for Selecting Items in Preschool Tests*, Minneapolis: University of Minnesota Press, 1946, Chapter 2.

²³ By A. R. Gilliland and A. M. Shotwell, Northwestern University, Evanston, Ill.

²⁴ See Cattell, *op. cit.*, 52 ff.; also N. Bayley, "Mental Growth in Young Children," *Thirty-ninth Yearbook*, National Society for the Study of Education, Bloomington, Ill.: Public School Publishing Co., 1940, Part II, pp. 11-47.

differences in the age of appearance of various functions and differences in their rates of development—appearance of new functions and changes in rates being especially rapid in the first two years of life.

Closely allied to the foregoing is the fact that tests included in infant scales are dissimilar to those used at later age levels, so that little correlation is to be expected. The reader will have observed that tests used in appraising an infant's development in the first eighteen months of life are largely of relatively simple motor activities and of sensory perception. These have never been found to correlate significantly with tests used at later age levels, which increasingly involve the higher and more complex mental functions. (See the chapter on definitions and nature of intelligence.) It may be that psychologists will not be able to devise infancy tests having greater predictive value; for it seems that those functions which are subsumed under the term "intelligence" do not reach a measurable magnitude until an age later than infancy. That is, intelligence, as psychologically understood and defined, does not emerge sufficiently during the earliest phase of development.

In conclusion, then, it may be said that when tests are used with children below the age of eighteen months, emphasis should be placed upon an analysis of performances and their evaluation of the child's *present* status. Thereafter, the scales increase in value for the purpose of predicting later mental level. Bayley,²⁵ after surveying her own and other researches, concluded that tests given between two and four years of age will predict eight- and nine-year intelligence test performance with moderate success ($r = .55$); while tests given at four years of age will predict eight- and nine-year performance much more satisfactorily ($r = .75$). These conclusions, however, were written before the publication of Cattell's scale and its validity data. It appears, therefore, that, as Cattell has shown, it is possible to devise scales of significantly higher predictive value for use with preschool children who are above eighteen months of age.²⁶

²⁵ N. Bayley, *op. cit.*, pp. 16 ff.

²⁶ It would be highly desirable to have more psychologists and institutions devote more time, energy, and money to research on the very important problems in this area, and perhaps to divert some of the time, energy, and money now being lavished on the study of personality and projective methods.

NONVERBAL GROUP SCALES OF MENTAL ABILITY

BEGINNINGS

The original Binet scale and its several revisions are administered to one person at a time; hence they are called individual scales. This is true, also, of the performance scales already described. Individual scales, obviously, are time-consuming and require that the examiner be highly skilled in administering them, in interpreting responses, and in evaluating the subject's behavior during the course of the examination. Impelled, perhaps, by the prevailing urge for "efficiency" and mass production, and by a desire to investigate large-scale problems, American psychologists undertook to develop tests which could be administered to a group of persons—large or small—all at one time.

World War I provided the occasion for the organization of the first group test. Prior to 1917, psychologists had been experimenting with test items and organization with a view to group examination. Shortly after the United States entered World War I, a psychological branch was formed in the army in order to develop and use group scales for the purpose of general classification of soldiers on the basis of mental ability, so far as the tests might measure that trait. A few other devices were developed for use in the army—e.g., trade tests—but in the army of World War I the work and contribution of psychologists were very largely in the testing of general ability.

Although the 1916 Stanford Revision and the Yerkes Point Scale

were employed to some extent, as well as an individual performance scale, the main task in the army was one of testing very large numbers of men in a short space of time. Consequently, the Army Alpha scale (verbal) and the Army Beta scale (nonverbal) were organized, both being group scales. These were actually the product of the contributions of individual psychologists, notably Arthur S. Otis, who pooled their experience, experimental results, and resources.

About 1,750,000 men were tested in the army of World War I. Though the scales were by no means highly satisfactory instruments, and though the men were very often examined under unfavorable conditions, the results obtained were of some assistance in the selection of men for advanced or special training, on the one hand, and of men of such inferior ability as to be unsuited for military training, on the other.

The use of psychological testing in the army of World War I had many outgrowths, some of which were, no doubt, unforeseen by the psychologists themselves. The data were reported and analyzed in a huge volume.¹ On the basis of these data, many periodical articles and books appeared on such subjects as racial and national differences in intelligence, geographic differences in intelligence within the United States, differences between occupational groups, relationship between educational status and intelligence, and the general intellectual level of the American adult. Not only were many of these data of doubtful validity but some of the interpretations and publications based upon them gave rise to serious misapprehensions in regard to the foregoing problems, which are loaded with social and educational implications. Another result of psychological testing in the army was the impetus it gave to the development of group tests for civilian purposes, notably in educational work at all levels, from kindergarten through university. Also, it set a precedent, for in World War II psychological testing was conducted on a vast scale in all departments of the armed forces.

The types of test materials included in the army group scales and in the numerous group scales subsequently developed were not all innovations. For example, tests of memory, sentence completion, free and controlled word association, arithmetic computation, vocabulary, classification of objects, and following directions had been in process

¹ Yerkes, R. M., editor, *Psychological Examining in the United States Army*, Memoirs of the National Academy of Sciences, Washington: Government Printing Office, 1921, Vol. 15.

of experimentation beginning within the last twenty-five years of the nineteenth century, in the United States and in several European countries.

CHARACTERISTICS OF GROUP TESTS OF MENTAL ABILITY

With a few exceptions, group tests—implicitly or explicitly—are constructed on the principle that intelligence is a general capacity and that it should be measured by means of sampling a variety of mental activities. Inspection of the scales shows, therefore, that they include in various combinations such items as following directions, arithmetical problems, practical judgment (in connection with “common-sense” problems), word meaning, disarranged sentences, completion of number series, completion of sentences, verbal analogies, information, mazes, three-dimensional visualization and counting of cubes, symbol-digit combinations, picture absurdities, picture arrangement, geometrical construction (“paper-form-board”), and geometric pattern analogies. Samples of each of these will be presented later when illustrative scales are described.

In most group scales, the items of each type (e.g., number series) are placed together in separate subtests or parts, beginning with the easiest and progressing by intervals—as nearly equal as may be achieved—to the most difficult. The principle involved here is this: by means of such an arrangement of items, every individual for whom the test is intended should be able to get some items correct and to proceed to a level of difficulty which represents his maximum in that particular type of mental activity.

Occasionally, however, it will be found that items in a scale are arranged in “spiral omnibus” fashion: that is, items of various types are presented in regular or irregular order instead of being grouped separately in subtests. Thus, there may be a sequence of this kind: one item each in number completion, arithmetical problem, vocabulary, information, analogies, etc.; then the types of items, increasing in difficulty, will be repeated in the same or in a different order.

Every group scale is standardized for a specified range of ages or school grades.² Thus the particular types of items used and the levels

² Standardization for a grade range is, in practice, tantamount to standardization for an age range because the scale must be adequate for the spread of ages ordinarily located within those grades. Furthermore, even when a scale is specified for certain school grades, it provides tables of norms for the various age groups.

of difficulty within a scale will depend upon the group for which it is intended. For instance, a group scale designed for children from kindergarten through the second grade will be almost entirely nonverbal in character, except for directions; one designed for pupils in the intermediate grades will include an increasingly larger portion of abstract and conceptual items (verbal and numerical); while tests of intelligence for high-school pupils and college freshmen are very largely, some entirely, of the verbal and numerical kind.

On many group scales, an individual's score is first obtained in terms of the number of points earned; that is, a raw score. From a table of norms, this score is converted into a mental age, from which an intelligence quotient is calculated. The manuals of some group scales provide, also, tables necessary to find an individual's percentile rank for his age or grade, or both. Other group scales dispense with mental ages and intelligence quotients, and give only percentile-rank equivalents for the range of point scores.

Group scales are scored more rigidly and more objectively than individually administered scales, such as the Stanford-Binet.³ In the former, the correct response, or responses, are supplied for each item so that they can be scored by clerks or machines. In the case of the Stanford-Binet and similar scales, while specimens of satisfactory and unsatisfactory responses are supplied, it is frequently necessary for the examiner himself to evaluate some responses and to decide whether or not credit should be given for them. This necessary exercise of judgment, however, does not invalidate the scales; for correlational studies have shown that there is very close agreement between experienced examiners as to the scoring of given responses.

Most group scales impose time limits for each of the several subtests, or parts. Whether this fact makes a scale a test of speed of response, solely or largely, or whether the scale measures "power" (level of difficulty the individual is capable of reaching) is a question to which answers have been provided by experiment. The imposition of time limits does not necessarily make a scale a test of speed of performance; the significance of the speed factor, in affecting the total score of a person, varies with the scale used.

³ This does not mean that the group scales are therefore better instruments. In fact, inflexible scoring is a disadvantage if the examiner's purpose is to study an individual clinically and to analyze test results qualitatively as well as quantitatively.

Some group scales are entirely nonverbal in content; others are entirely verbal; while still others combine the two types of items. In this chapter we shall describe several representative scales of the non-verbal variety.

PINTNER-CUNNINGHAM PRIMARY TEST⁴

This scale (having alternate forms, A and B) is intended for children in kindergarten, grade 1, and the first half of grade 2. It consists of seven subtests: common observation (identifying objects commonly found in the usual environment); perception of esthetic differences; identification of associated objects (knowledge of relationships between objects, such as key and lock); discrimination of size; perception of the elements that constitute a whole picture; picture completion (noting missing parts in a picture and, from a series of choices, indicating the missing part); copying designs (using a given square of dots).

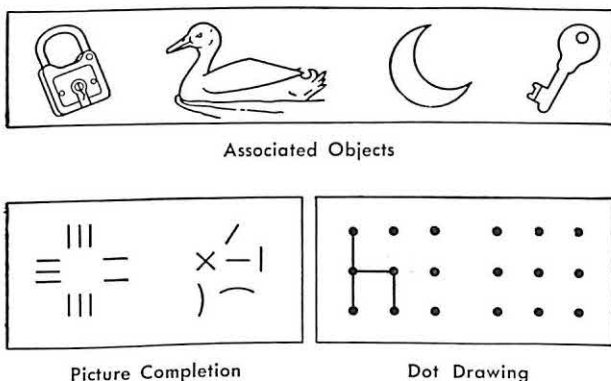


FIG. 10.1. *Items from Pintner-Cunningham Primary Test.* Copyright 1938 by the World Book Company. (Reproduced by special permission.)

Mental age norms are provided from 4 years and 1 month to 10 years.

Validity and Reliability. As in the case of many other group scales, validity of the Pintner-Cunningham test is given in terms of its cor-

⁴ By R. Pintner, B. V. Cunningham, and W. N. Durost. Published by World Book Company, 1938.

relation with the Stanford-Binet (1916), the reported coefficients ranging from .73 to .88.

Reliability is reported in terms of correlations between the alternate forms (A and B) and in terms of the probable error of test scores. The reliability coefficients vary from .83 to .94. The standard deviation of scores is reported as being about five times as large as its probable error; hence this criterion of reliability may be considered satisfactory, since the minimum acceptable relationship between a standard deviation and its probable error is usually given as three to one.⁵

CHICAGO NONVERBAL EXAMINATION⁶

This scale is designed for subjects from six years of age through adulthood;⁷ but it is very doubtful whether it is at all adequate for a representative sampling of persons above age thirteen. The authors state that the scale has proved clinically useful for children between seven and thirteen years of age. In the case of persons above age thirteen, it appears that the tests measure speed of performance.

The types of items included are not unique: symbol-digit; perception of similarities and classification of objects (by crossing out the one object that is of a different class); three-dimensional visualization and block counting; "paper-form-board" (marking the parts that would make up a whole geometric figure); visual perception of detail (matching geometric designs which have more or less internal detail); picture arrangement (numbering parts of wholes, to indicate how the parts must be placed to form the whole); numbering pictures in their correct order to represent a logical sequence of events; picture absurdities (noting the missing or superfluous parts); picture matching by relating a part to a whole picture; symbol digit (more complex and more extensive than the first symbol-digit test).

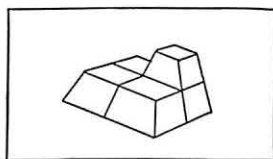
Validity and Reliability. The reliability coefficients reported are between .80 and .90. Some of these coefficients are within the range usually regarded as satisfactory. But they were calculated for age

⁵ The probable error of the standard deviation indicates the extent of fluctuations in scores to be expected as a result of random errors of measurement.

⁶ By A. W. Brown. Published by The Psychological Corporation, 1936.

⁷ The scale may be administered by using verbal directions or by means of pantomime. Pantomime, however, cannot be used successfully with children younger than eight years.

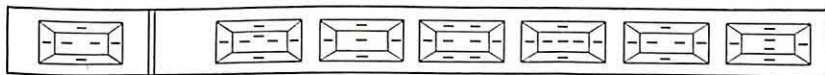
ranges of two and three years and extend over two to six school grades, whereas reliability coefficients for each chronological age group and each school grade would be more valuable, if a child's performance is to be compared with others of his own age or grade.



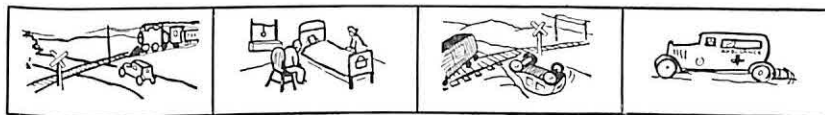
Block Counting



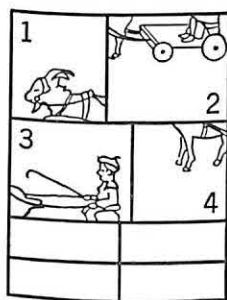
"Paper Form Board"



Matching Figures



Picture Sequence



Picture Arrangement

FIG. 10.2. Items from *Chicago Non-Verbal Examination*. Psychological Corporation. (By permission.)

The criteria of validity used in standardizing this scale were the usual ones: (1) correlation with CA, (2) known groups (in this instance, the mentally retarded), (3) symmetrical bell-shaped distribution, and (4) correlation with results of other tests.

The reported coefficients of correlation for test scores and chrono-

logical age varied from .57 to .81, using groups having an age range of seven or eight years.

Employing the second criterion, the authors of the scale found an average IQ of 61 (S.D. = 12) for 99 feeble-minded children, as compared with an average of 62 (S.D. = 6) obtained with the Stanford-Binet. The mean difference between the Stanford-Binet and the Chicago Nonverbal ratings, however, was 9.0 points (disregarding signs; correlation coefficients were not given). Thus, while the two means are practically identical, the standard deviations and the mean of the differences are such as to indicate that in many individual instances there were significant discrepancies between ratings obtained on the two scales. These discrepancies are most probably attributable to the differences in content of the scales; and they suggest, further, that the two scales may be used to supplement each other, rather than as substitutes, when an individual's mental abilities are being analyzed and evaluated.

In respect to the third criterion, the authors report a satisfactory distribution of scores, closely approximating the symmetrical curve. Like Terman and many other psychologists, they believe that measured intelligence should be and is naturally distributed in the form of the "normal frequency surface"; hence they use that curve as a criterion.

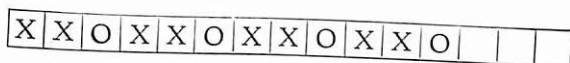
Interestingly enough, excepting the mentally deficient, the results of the Chicago Nonverbal Examination have been compared with results of two other group tests that are largely verbal in content (Otis and Kuhlmann-Anderson), but not with those of the Stanford-Binet which, by most specialists, is considered to be a most valuable criterion in standardizing group tests. With chronological age not held constant, the coefficients fell between .57 and .74; but with CA constant, the coefficient was only .51. These correlations represent only modern correspondence.

REVISED ARMY BETA EXAMINATION⁸

The original Army Beta scale was constructed in order to provide an examination for illiterates and non-English-speaking men in the Army of World War I. It included the familiar mazes, cube analysis and counting, symbol-digit combinations, pictorial completion, and geometric construction. There are two other subtests with

⁸ By C. E. Kellogg and N. W. Morton. Published by The Psychological Corporation, 1935.

which the reader may not yet be familiar: the first is the X-O series, which includes a number of arrangements of the letters X and O; each series has a number of blanks that are to be filled in according to the arrangement of the given sequence. For example:



The other is a number-checking test, which consists of a series of paired numbers, from short to long. The subject is required to check those pairs which are identical. For example:

650
659012534

650
659021354

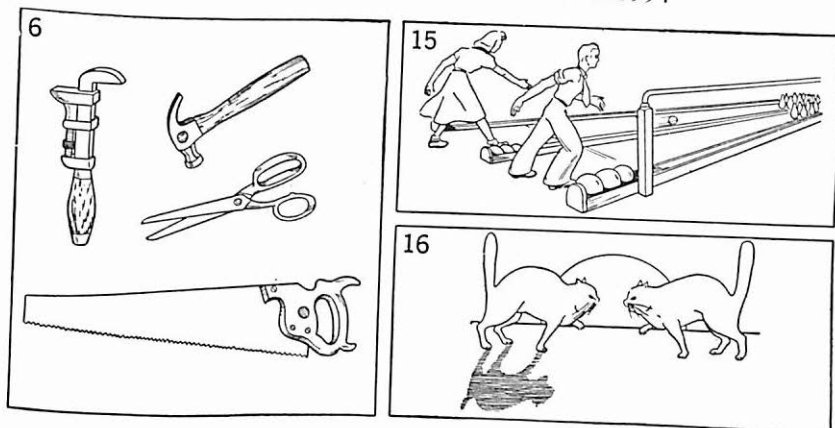


FIG. 10.3. Items from Revised Army Beta Examination. Psychological Corporation. (By permission.) For No. 6 the testee marks the "picture absurdity"; for Nos. 15 and 16 he indicates the "missing parts."

The Revised Beta Examination, also, is intended to serve as a measure of general ability in the case of "relatively" illiterate or non-English-speaking subjects. The types of subtests of the two scales—original and revised—are not identical; for in the revised scale, the cube counting and the X-O tests have been omitted, while picture absurdities and object similarities and differences have been added.

Norms are provided in terms of Stanford-Binet (1916) mental age equivalents, the range being from MA 6 years and 3 months to 16 years and 8 months.

Validity and Reliability. The scale's reliability coefficients were high when odd-even items were correlated, r being .987. But the coefficient for test-retest scores was .77.

Validity of the revised Beta scale is given in terms of the correlation between its point scores and Stanford-Binet (1916) mental ages, the coefficient being $.78 \pm .02$. When the revised Beta point scores were correlated with the Otis Self-Administering Tests of Mental Ability (higher examination A), the coefficient found was $.71 \pm .02$. Since these coefficients are based upon results obtained with groups of subjects whose ages varied, the correlations must be interpreted with due regard for the fact that they would be lower if chronological age had been held constant. What this signifies, then, is that this non-verbal scale may be used to supplement verbal scales; but it should not be used as a substitute for them.

PINTNER NONLANGUAGE SERIES: INTERMEDIATE TEST⁹

This scale, devised for use from grade 4 through grade 9, differs from most other group scales of general ability in that it utilizes no verbal situations and is independent of word knowledge and language facility except as these are involved in understanding directions. The author also provides directions for administering the scale in pantomime where this is desirable, as in the case of subjects who suffer from language handicaps or from defective hearing.

The tests consist entirely of materials of a diagrammatic nature, intended to provide "relatively independent" measures of the "spatial factor," "perceptual ability" (visual), and "reasoning" (without use of language). This Pintner scale, therefore, is one of the few which specifically utilize some of the "factors" presumably isolated by those adhering to the group-factor theory of intelligence. Yet, interestingly enough, though tables for separate standard score ratings are provided for each of the subtests, no attempt is made to indicate which of the presumed factors is being measured by means of each of the six subtests. In fact, the author states that, "No claim is made that the subtests tap primary or independent abilities, and little is known as to the significance of the separate subtest scores. . . . Only large deviations [from the median] should be given any credence in guidance of individuals."¹⁰ Apparently, then, the inference that this

⁹ By R. Pintner. Published by World Book Company, 1945.

¹⁰ *Manual of Directions*, pages 1 and 9.

scale measures the spatial factor, perceptual ability, and reasoning rests entirely on an *a priori* basis rather than upon scientific analysis.

The total scores of the scale, however, are regarded as having validity; for tables of norms, from age 7 to age 17 are provided, from which mental ages, intelligence quotients, and percentile ranks may be obtained.

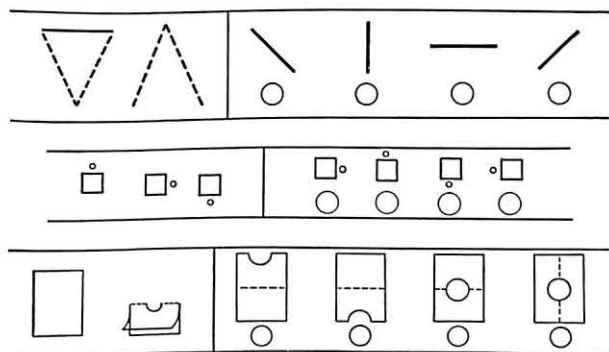


FIG. 10.4. Items from Pintner General Ability Test—Intermediate. Copyright 1938 by World Book Company. (Reproduced by special permission.) At top are Reverse drawings: after looking at the sample pair to the left of the bar, the subject looks at the first drawing to the right of the bar and finds its reverse among the next three drawings. For the right middle series, Movement sequence, he finds which of the four drawings at the right completes the series started at the left of the bar. For the bottom series, Paper folding, he indicates which of the items at the right of the bar shows the way the folded paper at the left of the bar would look if it were opened up.

Validity and Reliability. Reliability of the scale is reported in the usual terms of correlation coefficients between scores on odd-numbered items and scores on even-numbered items, these being .85 (Form K) and .89 (Form L).

As criteria of validity of his nonlanguage scale, Pintner used increase of mean scores in successive age-groups, correlation with his verbal tests of ability, and similarity (judged by observation) between his tests and those devised by factorial analysts. The increases in mean scores are moderate; and the correlation coefficients (second criterion) are in the .60's, for groups whose age range is only one

ear. Although these coefficients indicate a significant amount of relationship between the verbal and nonverbal tests, they are still far enough removed from unity to warrant again the conclusion that these two types of tests may not be used interchangeably but can be used to provide a fuller understanding of an individual's abilities, particularly in those instances where a person is handicapped by language disability, or where it is desired to discover the existence of a specific type of disability (e.g., in visual perception and discrimination) of a nonverbal kind in the case of a person who has no language handicap.

NONLANGUAGE MULTI-MENTAL TEST ¹¹

This scale (in two forms, A and B) employs only a single type of item throughout. Each item consists of drawings of five objects, four of which "belong" on the basis of some common relationship, while the fifth does not "belong." The testee has to identify and mark the non-belonging item.

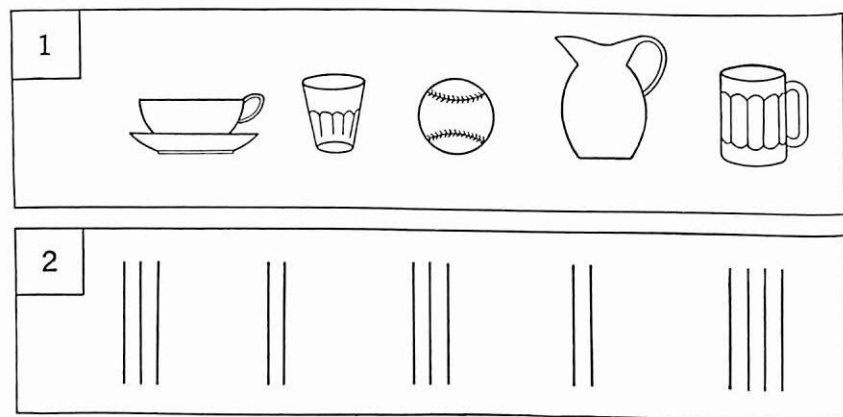


FIG. 10.5. Recognition of Relationships. Items from Non-Language Multi-Mental Test, Teachers College, Columbia University. (By permission.)

The authors state that their scale ". . . is constructed to measure the ability to recognize and utilize relationships not among verbal symbols, but rather among pictorial symbols."¹²

Norms are provided for mental-age equivalents and grade equivalents.

¹¹ By E. L. Terman, W. A. McCall, and I. Lorge. Published by Bureau of Publications, Teachers College, Columbia University, 1942.

¹² *Manual of Directions*, p. 1.

lents; in the former instance from a mental age of 33 months to one of 236 months; in the latter instance, from mid-first grade to mid-eighth.

Validity and Reliability. Two common criteria of validity were used: correlation with chronological ages ($r = .65$) and correlation with mental ages as derived from a variety of group tests of intelligence ($r = .67$).

Reliability coefficients of the scales, for a grade range of 3 to 8, were .86 for Form A, .90 for Form B, and .94 for both forms combined. Reliability coefficients for each of the several grades taken separately varied from .66 to .74, for each form taken alone. When both forms were combined, estimated reliability was .80.

PATTERN PERCEPTION TEST¹³

This test, like the one immediately preceding, employs a single type of material. But it is much more complex, its solutions require more subtle perceptions and reasoning, and it is intended for use only with adults.

The scale includes sixty-four items (or problems). Each item consists of a row of five designs, the problem being to discern the four designs which form a pattern and to cross out the extra or inappropriate one. There are eight sets of items, each set placed in the order of increasing difficulty. In each set of items the problems begin with an elementary presentation of a theme, or pattern, which is developed with increasing complexity in subsequent items.

Although the Pattern Perception Test is still in process of standardization and development, it is described here because it represents a type of nonverbal material that may prove to be very valuable for use

¹³ Prepared under the direction of L. S. Penrose. Published by Galton Laboratory, University of London, 1947. This type of test was originally planned by Penrose and Raven; later Raven standardized a form principally for individual testing. During World War II, a short form for group use was prepared by Raven for use in the British army. See L. S. Penrose, "An Economical Method of Presenting Matrix Intelligence Tests," *British Journal of Medical Psychology*, Vol. 20, Part 2, 1944, pp. 144-146. For background of the present test, see L. S. Penrose and J. C. Raven, "A New Series of Perceptual Tests, Preliminary Communication," *ibid.*, Vol. 16, 1936, pp. 97-104; J. C. Raven, "The R.E.C.I. Series of Perceptual Tests: an Experimental Survey," *ibid.*, Vol. 18, 1939, pp. 16-34; P. E. Vernon, "Research on Personnel Selection in the Royal Navy and the British Army," *The American Psychologist*, Vol. 2, 1947, pp. 35-51.

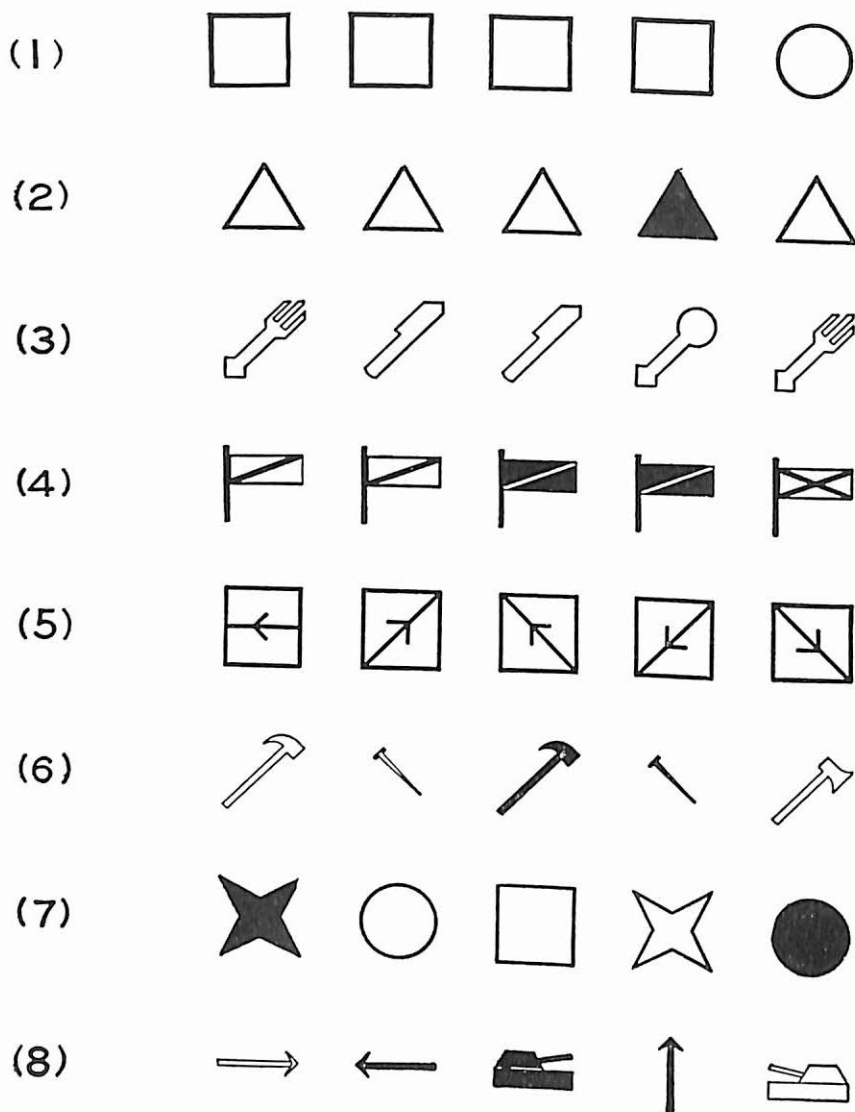


FIG. 10.6. Items from Penrose Pattern Perception Test. Galton Laboratory, University of London. (By permission.)

with a wide range of adult ability. This test has been found, for example, to have value in identifying men at both the lower and the upper ends of the distribution of mental ability. It may prove to be valuable, also, in examining psychotic persons (or those suspected of

being psychotic); for reports indicate that these persons perform relatively poorly on items requiring judgment (insight) and constructive thinking.

Statistical data thus far reported show test-retest reliabilities to vary between approximately .80 and .90. Validity coefficients determined by correlating the test scores with job ratings in the British navy and army varied widely, due to the degrees of reliability of rating criteria and the effects of selectivity in various jobs. The *mean* validity coefficients for each of a number of navy and army branches varied from .30 to .47. Correlations with other standardized tests range from a low of .43, for 67 medical students (a fairly homogeneous group) to .73, for a random sample of 597 men in the British army.

PROGRESSIVE MATRICES TESTS ¹⁴

These are nonverbal scales designed to evaluate the subject's ability to apprehend relationships between geometric figures and designs, to perceive the structure of the matrix and of the figure (part) necessary to complete each system of relations (the matrix) presented. The tests, thus, are intended to evaluate the person's ability to discern and utilize a logical relationship presented by these nonverbal materials. The problems require, in varying degrees, analytical and integrating operations of the kind called "insight through visual survey." Verbalization and abstraction of relationships are also possible factors, if the subject is able to analyze and synthesize by these means. Factorial analysis suggests that the matrices tests are measures largely of a "general factor," with a small loading of a spatial perception factor. Raven, the author of these tests, interprets this factor as being essentially the same as Spearman's eduction of relations and eduction of correlates.

There are several sets and editions of the matrices tests, each intended for a specified age group or for limited ability levels. Thus, one set is for children between the ages of 3 and 10, and for mental de-

¹⁴ Prepared by J. C. Raven. Published by H. K. Lewis & Co., Ltd., London. See by Raven, "Standardization of Progressive Matrices, 1938," *British Journal of Medical Psychology*, Vol. 19, Part I, 1941, pp. 137-150; "Progressive Matrices," published by The Crichton Royal, Dumfries, Scotland (1938-1951). Also, G. Keir, "The Progressive Matrices as Applied to School Children," *British Journal of Psychology* (Statistical Section), Vol. 2, 1949, pp. 140-150; G. A. Foulds and J. C. Raven, "An Experimental Survey with Progressive Matrices (1947)," *British Journal of Educational Psychology*, Vol. 20, 1950, pp. 104-110.

fectives; another may be used with children from age 6 and with adults; a third is devised for use with only the highest quarter of the population.

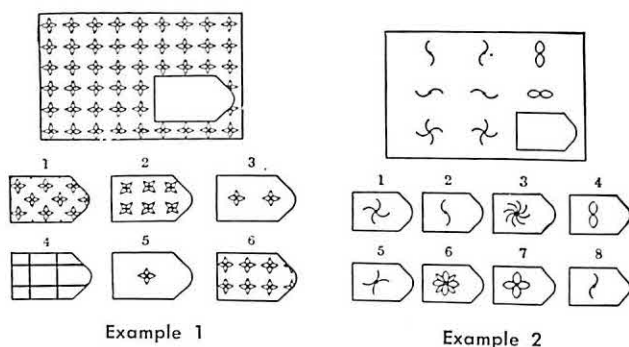


FIG. 10.7. Specimen Items from Raven Progressive Matrices Test. (By permission.)

Validity and Reliability. While a fair amount of research has been carried out regarding reliability and validity, much more remains to be done in these respects. Reliability has been found to differ at the several age levels; and while validity has been suggested in terms of a general factor, the test's predictive efficiency—that is, their practical validity—needs further investigation. Even so, this type of instrument is a very valuable addition to available nonverbal tests, particularly since it shows considerable promise for use with adults at the superior as well as average levels. Progressive matrices, as a type of test material, and the pattern perception type also, merit intensive experimental study, in themselves, and experimental comparison with the types of materials commonly included in nonverbal tests developed and widely used in the United States. At the present time, both of these instruments have progressed sufficiently to warrant their use as valuable supplements to the scales currently employed in both nonclinical and clinical situations; for these tests have been found useful with and interesting to individuals of a wide variety of ages, ability levels, and degrees of stability (and instability).

CATTELL CULTURE-FREE TEST

This is an attempt to provide a measure of general mental ability free from verbal materials and from "the acquired skills of

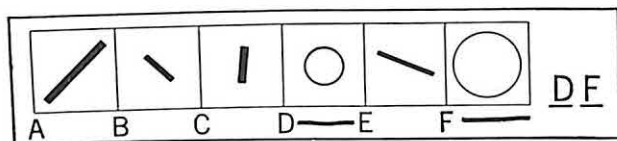
most performance tests.”¹⁵ It consists of six parts. The first, a classifications test, requires the subject to identify in each row of six figures the two that do not belong with the others. In the second, called “pool reflections,” the subject identifies the one of six drawings which represents the specimen drawing as it would appear in a pool image. This is a test of spatial perception. The third is a completion test in which the testee identifies from among six specimens the one drawing that will complete a series of four members. Completions vary from simple matching to the eduction of correlates. The fourth, fifth, and sixth parts are called *matrices*. In these, the subject has to identify the last member of a series of four or nine parts, the patterns being of increasing complexity. Some of the sequences are horizontal; some are vertical; some involve rhythms or cycles. The last of the subtests, the sixth, is further complicated by the fact that several sections of the matrix are missing, thus making it more difficult to discern the pattern.

Validity and Reliability. Reliability is indicated by a split-half coefficient of .88 (corrected by the Spearman-Brown formula), obtained with a group of 121 high-school freshmen.

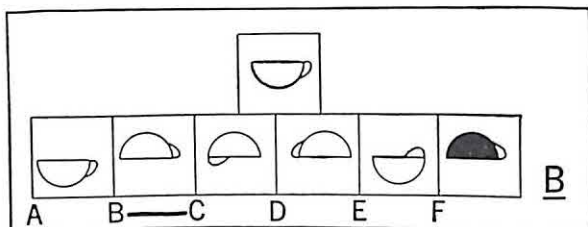
To begin with, this scale was standardized on about one hundred boys in a junior vocational high school and the same number in an academic senior high school. In retaining or rejecting an item, the criterion was its ability to discriminate between the two groups, the academic school pupils being regarded as the more able. (Note the assumption here that vocational school pupils are mentally less able than the other group. The differentials, presumably, are not regarded as due to the age differences.) The retained items were further screened on the basis of responses obtained from college students and from pupils in grades seven and eight. Finally, an analysis of items was made on the basis of highest-scoring and lowest-scoring individuals in a group of two hundred students doing major work in psychology.

Although this is called a test of general ability, there is a low communality of functions measured, as judged from the intercorrelations of the parts. These run from .12 to .63, with a median of .38. The correlations of each of the part scores with total scores varied from .55 to .82, the median being .74. These coefficients, however, are spuriously high because the score of each subtest was included in

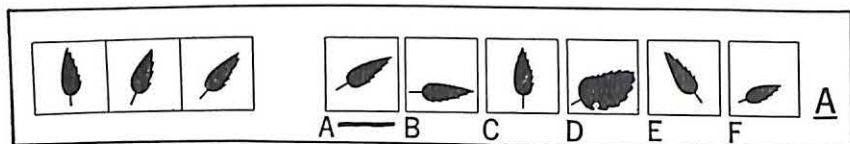
¹⁵ By R. B. Cattell. Published by The Psychological Corporation, 1944.



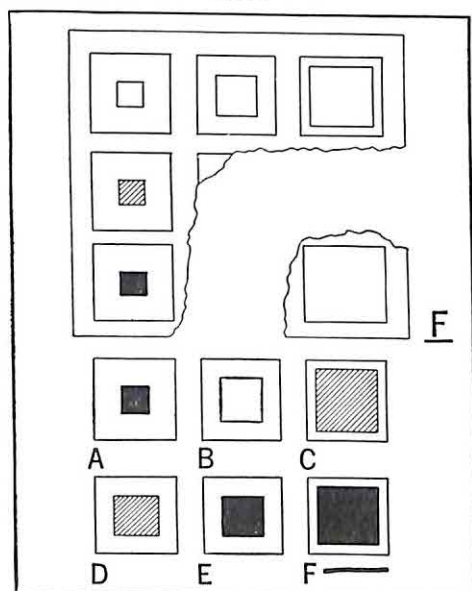
Classification



Pool Reflections



Series



Matrices

FIG. 10.8. Items from Cattell Culture-Free Test. Psychological Corporation. (By permission.)

the total score; the coefficients, therefore, are in part indicative of self-correlation.

When results of the Culture-Free Test were correlated with scores on the Modified Alpha Examination, Form 9, the median coefficients for four groups ($N = 376$) were about .50 with Alpha verbal and numerical scores taken separately, and about .55 with the total Alpha scores. The Culture-Free Test was also correlated with the Stanford-Binet, the American Council, and the Arthur scales, and with the Ferguson Form Board results.¹⁶ The mean coefficient with these was .52.

The foregoing coefficients would be considered too low as evidence of validity, if the other intelligence tests were taken as basic criteria. Cattell and his collaborators apparently do not do so; they offer these data, it seems, to indicate the extent of agreement between their test and the others. Basically, however, their criterion is one of factor validity based upon acceptance of the general-factor theory and upon factor analysis of their data. Their view is essentially this: Devise a sampling of tests which have been accepted by psychologists as testing intelligence—in this instance nonverbal perception of relationships and spatial perception; that is, those tests demanding insight into complex relations, ability to learn, ability to think abstractly, and ability to cope with new situations. Analyze the test results to find if they yield a general factor. This was done; statistical evidence of a general factor was found; and Cattell concluded that he and his collaborators were measuring general ability.

None of the pragmatic and usual criteria of validity were applied, excepting correlations with other tests, but to these not much significance was attached. This Culture-Free Test should be regarded as a tentative and experimental device until its validity is demonstrated in terms other than factorial analysis.

GOODENOUGH DRAWING TEST¹⁷

This is a test which purports to evaluate a child's intelligence by means of his drawing of a man. It is intended for ages 3½ years to 13½. The child is instructed to make a picture of a man as best he

¹⁶ R. B. Cattell, S. N. Feingold, and S. B. Sarason, "A Culture-Free Test of Intelligence: II. Evaluation of Cultural Influence on Test Performance," *Journal of Educational Psychology*, Vol. 32, 1941, pp. 81-100.

¹⁷ F. Goodenough, *Measurement of Intelligence by Drawings*, Yonkers, N. Y.: World Book, 1926.

can. He is told to work carefully and to take his time. Scoring is based not upon esthetic quality but, rather, upon the presence of essential details which presumably indicate the individual's level of perceptual differentiation of an object that is very familiar in his environment.

Although the reliability coefficients are often satisfactory, ranging from about .75 to about .90, the drawing test does not correlate well with the other types of tests which have been found most useful and promising in the measurement and evaluation of general ability. Yet, this test is another instance of a device which, in spite of its doubtful general validity as a test of intelligence, has been found useful as an adjunct to verbal tests when mental deficiency is suspected in the case of a given child.

DAVIS-EELLS TEST OF GENERAL INTELLIGENCE¹⁸

Rationale. This test of "problem solving ability" is an important innovation in the testing of intelligence. It is unusual in respect to its content and the rationale thereof, and in respect to its frank rejection of statistical validation with other tests. In place of such validation, the authors substituted intensive interviews with children to disclose the "mental problems of a kind found in most of the basic areas of children's lives: school, home, play, stories, and work. The specific problems resulted from intensive observation and detailed interviewing of children in many areas of activity. . . . The extent to which the items in the test deal with problem situations which seem real to children of the age levels for which the test is planned may best be judged by examining the test with this criterion in mind." The test items selected for initial tryouts and experimentation were based upon the "insights of a number of educators and sociologists familiar with the characteristic modes of living and child upbringing at different socio-economic levels, and in part upon systematic observation of children in free-time activities (on playgrounds, in neighborhood groups, in schools, etc.)."¹⁹

Contents. The instrument finally emerged after extensive research and six tryouts with large numbers of children, from widely different socio-economic backgrounds, in several sections of the United States. It is believed to be culturally fair to all socio-economic groups in *urban* areas.

¹⁸ By Allison Davis and Kenneth Eells. Published by World Book Co., 1953.

¹⁹ *Manual*, pp. 7 and 18.

Forms for two levels are available: Primary, for grades 1 and 2; Elementary, for grades 3 through 6. The items finally selected are 88 in number, as follows:

Best-Way items (29). In each item, three pictures show the beginnings of attempts to solve a stated problem or perform a given

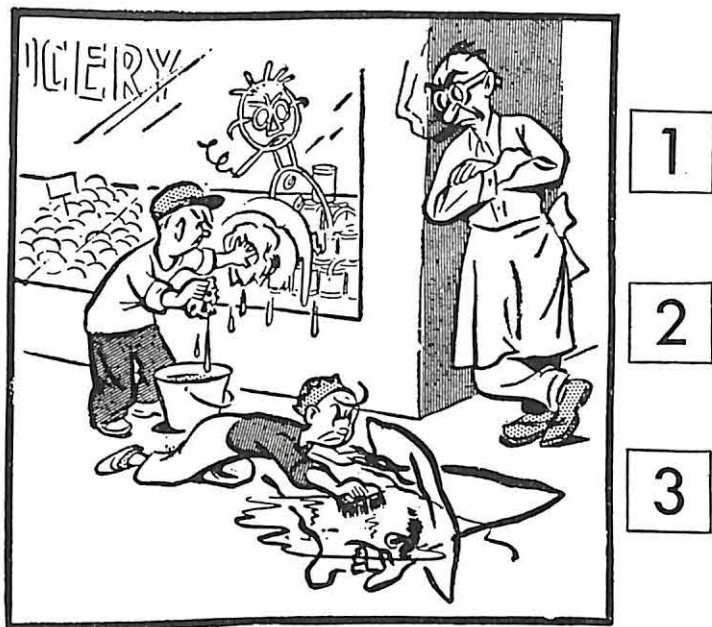


FIG. 10.9. The task is to select the one statement that best explains the situation shown in the picture.

1. The boys want to wash the man's window and sidewalk.
2. The man is making the boys wash his window and sidewalk.
3. You can't tell from this picture why the boys are washing the window and sidewalk.

From Davis and Eells Test of General Intelligence. World Book Co. (By permission.)

task. The subject indicates the picture that will lead to the best solution of the problem.

Probabilities items (29). Each picture shows a situation in which certain elements are present. The child has to select, from three choices, the most probable explanation of what is happening in the picture.

Picture Analogy items (22). This is the familiar type of item in which a relationship between two objects is shown, and the subject is required to find a similar relationship in a given set of pictures.

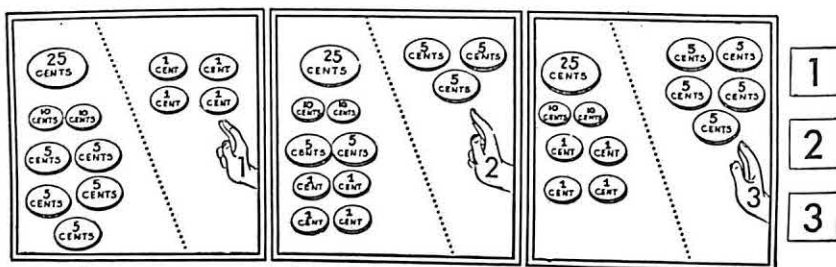


FIG. 10.10. The task is to select the picture in which a given sum of money can be made, beginning with the right-hand side of the dotted line and completing from the left side. In this problem, the sum to be made is 40 cents. From Davis and Eells, op. cit. World Book Co. (By permission.)

Money items (8). In each item, two sets of coins are shown in three different combinations. Each combination is incomplete. The problem is to discern the appropriate combination, of the three, that will yield a stated sum when completed.



FIG. 10.11. The task is to place the bottles in the black box in such a way that the white box may best be placed on the black one. From Davis and Eells, op. cit. World Book Co. (By permission.)

Of these, 26 are in the Primary form only, 41 are in the Elementary only, and 21 are common to both.

Validity and Reliability. Although Davis and Eells do not base the validity of their tests upon correlations with other and earlier scales, they do present, for informational purposes, some correlations with

the Otis Quick-Scoring tests, obtained for groups in single grades, 3 through 6. Of the sixteen coefficients, seven are in the .50's; the remainder are rather evenly distributed from .39 (lowest) to .66 (highest). The authors believe these coefficients are what should be expected, since they indicate that the abilities measured by their tests bear a substantial relationship to those measured by the other tests; yet their tests and the others do not measure altogether the same factors.

Correlations with standardized school achievement tests are also reported. These are:

with reading:	.43
with arithmetic:	.41
with language:	.40
with spelling:	.24

These coefficients are significantly lower than those found with the more usual types of individual and group scales; but this is to be expected because of the very nature of the Davis-Eells tests, and since the authors' position is that several significant factors other than problem-solving ability contribute to success in school achievement.

Split-half reliability coefficients range as shown below:

grade 1:	.68
grade 2:	.82
grade 3:	.84
grade 4:	.83
grade 5:	.82
grade 6:	.81

A coefficient of .68 is much too low for predictive purposes. The remaining indexes are moderate, but not as high as is optimally desirable.

Test-retest reliabilities (two-week interval) are:

grade 2:	.72
grade 4:	.90

On the whole, these coefficients, if typical, indicate that the present test is not so reliable for refined differentiation as it is for differentiating between levels; for example, very superior, superior, average, inferior, very inferior.

In connection with the validity and general characteristics of these tests, several aspects should be pointed out. The authors aimed to develop an instrument that would present situations and problems in forms that are within the experiences common to all children in the groups for whom they were planned. This meant the elimination of language (except in the directions) and other cultural factors that might favor one group and handicap another. While nonverbal tests are not new, the kinds of situations presented are novel. Furthermore, the psychological functions to be sampled by these and similar problem-situations were determined and specified after psychological interviews and psychological analysis, rather than by factorial analysis.

Evaluation. The foregoing difference in procedure is very significant with regard to principles of test construction. Factorial analysts devise a large number of test items and subtests on the basis of certain assumed mental processes involved; the scores on these are intercorrelated and further analyzed in order to find common functions, to combine and reduce the number of subtest categories, and to name them. Davis and Eells, on the other hand, first ascertained the kinds and range of problems children deal with: then by interview and analysis of children's responses they determined that certain psychological functions were operative and had differentiating significance. The functions they specify are association, insight, reasoning, and organizational ability (method of attacking a mental problem). After the items for these tests were developed, each of a group of children was interviewed to determine whether the problems evoked the mental processes which the tests seek to measure. It was found that 92 percent of the children who answered analogy problems correctly also explained the analogy-relationship correctly. And nearly all pupils explained the relationships correctly in solving the other types of subtest problems. These findings strongly indicate that the test problems evoked in the successful subjects the mental processes which the authors intended should be utilized.

Like all other new tests, the Davis-Eells will have to be subjected to investigation to determine their predictive efficiency. Since the authors point out that school achievement is dependent only in part upon problem-solving ability (which is to be measured by their tests), it is to be expected that predictive efficiency will not be very high, in terms of correlations, with school grades. It will be necessary, then, to

use other selected criteria of demonstrated problem-solving ability; or to utilize pupils in whom this ability is the only significant variable, while the others are practically constant.

EVALUATION OF NONVERBAL GROUP SCALES

Uses. A survey of available nonverbal scales shows, for the most part, that they are valuable with children who have had limited educational opportunities or impoverished social backgrounds, with young children who have not yet learned to read, with older pupils who are handicapped by reading or language difficulties, and with illiterate or non-English-speaking adults. Possible exceptions to this statement of limited usefulness are the Pattern Perception, the Progressive Matrices, and the Davis-Eells Tests.

Nonverbal tests are valuable, also, for the better diagnosis of cases who, on verbal tests, have intelligence quotients between about 60 and 75, and who, therefore, would be considered as subjects for special educational treatment or possibly institutional care. The examining psychologist might be in doubt with regard to such borderline cases; but if the results of the nonverbal tests confirm those of the verbal, he has reason to allay his doubts. However, if the rating on the nonverbal tests is significantly higher, then the case will require further study to account for the discrepancy.

Nonverbal tests can be clinically useful, also, with individuals whose intelligence quotients are higher than 75; that is, for individuals who, on verbal tests, *appear* to be significantly less capable than there is reason to believe they actually are, on the basis of other information about them. In this connection, they are particularly useful in population centers having large numbers of non-English-speaking homes.

In whatever situation nonverbal tests are used, the examiner must realize that defective vision or slow psychomotor responses can be a handicap. The first of these handicaps points up the importance of clear drawings—a condition that is not always satisfied.

Tests of mental ability have had their greatest usefulness in schools, where they have been utilized for purposes of educational and vocational guidance, as well as in the diagnosis of learning difficulties in the case of a particular individual. Nonverbal group tests have been found valuable in efforts to determine aptitude and promise in shop work, mechanical drawing, architectural drafting, and occupations of a mechanical or quasi-mechanical nature—all of which make demands

upon those psychological activities which enter into problems involving geometric perceptions and reasoning with the concrete rather than with the abstract.

For these purposes, nonverbal group scales may not be used indiscriminately, for the following reasons. Some are measures principally of detailed visual perception (Revised Army Beta); some attempt to span too wide an age range and are not able to make any but rather gross differentiations (Non-language Multi-Mental Test, which is scaled for ages 33 months to 236 months); at the higher age levels, some are less suitable for women than for men because of specialized content.

Validity. Studies of the validity of nonverbal scales show that while most of them correlate significantly with scales of the verbal type (individual and group), the coefficients are far enough removed from unity to warrant using the two types as supplements rather than as equivalents. When scores on verbal and on nonverbal scales are correlated, for children in the earlier grades (approximately through grade six), the coefficients obtained are usually in the .60's and .70's, with relatively few in the .80's. But when the subjects tested are pupils in the later grades, the coefficients usually fall in the .50's and .40's, with a few lower and a few higher. These generally lower coefficients, in the case of pupils in the later grades, are due to the inability of most available nonverbal tests to discriminate between individuals in the upper levels of ability.

Functions Measured. The reader will recall that most, though not all, authors of nonverbal tests of mental ability seek to measure the same mental processes as those tested by means of verbal scales. Some of these authors are unequivocal in maintaining that the nonverbal tests require essentially the same type of intelligent performance as that required by the abstract symbols of language and number. They hold that the problems presented in diagrams, pictures, charts, and geometric forms closely parallel those presented by means of language and number. For example: picture arrangement is regarded as being similar in function to disarranged sentences; picture analogies similar to word analogies; picture completion similar to sentence completion; reasoning with geometric patterns similar to reasoning with numbers and words; perceiving similarities, differences, and part-

whole relationships in pictures and patterns similar to such relationships in language. Many nonverbal tests, however, suffer from the fact that they attempt to assess general ability by means of homogeneous test items or by means of a very limited number of subtests.

Related to the question of measuring general ability by means of a more or less homogeneous nonverbal scale, it is important to note that it is quite possible to find *some sort* of general factor—which is actually a statistical and theoretical concept—in a series of subtests such as those in Cattell's and others, but the statistical derivation of a general factor does not in itself give proof or assurance that the tests are necessarily measuring the functions being attributed to them. For example, a series of tests might be devised in which a "general factor" emerges, and which upon *psychological* analysis proves to be a factor of speed of work. In another instance a general factor of visual acuity may emerge, etc. Fundamentally, in any statistical analysis of test data, the factor or factors that emerge will depend upon the initial characteristics of the tests themselves. In establishing validity of a scale, therefore, it is inescapably necessary to analyze the psychological processes involved in the tests and to compare the results they yield with forms of activity and evidences of ability which are indicative of intelligent behavior in our culture.

The significant coefficients of correlation found between verbal and nonverbal tests of intelligence demonstrate that there is merit in the view that the two types are, in a significant degree, measuring the same or associated functions. But this does not mean that verbal and nonverbal tests are equivalent; for one type may also involve certain functions not involved in the other, or one may demand a higher level of the same functions being tested than does the other.

Language and number are symbolic systems which *represent* something else: e.g., objects, qualities, events, actions. Development of abilities in language and number facilitates intelligent behavior; for the use of these symbols expands the individual's range of experience beyond the limits of the immediate situation. Development of language and number makes possible a finer discernment of forms and objects in the world surrounding the individual; for with the use of language and number he is enabled to analyze, synthesize, classify, and organize his perceptions. Objects and events, at first vague, are more sharply defined; likenesses and differences are accentuated; evaluations are refined. Language and number enable individuals

also to organize their thinking into larger and more comprehensive unified patterns.

It is because the use of language and number requires that the individual go beyond the immediate concrete situation and because he thereby can engage in more complex and subtle mental operations that many psychologists regard ability to deal with symbols as a higher form of intellectual activity than the ability to deal with concrete objects. They therefore prefer to test intelligence, whenever possible and appropriate, by means of verbal and numerical materials. They would, however, use nonverbal tests, when these are made necessary by developmental immaturity, language, or cultural handicap, to gain the insights that these tests provide if they are adequately scaled in difficulty.

Cultural Influences. The emphasis upon verbal and quantitative aspects of intelligence in many of the individual and group scales has given rise to a misapprehension regarding the nonverbal scales: namely, that these latter are "culture free"; and, in fact, one test author (Cattell) has so named his scale. Inspection of the items in this scale and others reveals that they utilize many objects that children and older persons learn about through experiences in their environments. These experiences are dependent upon a culture just as development of verbal and quantitative abilities are. The differences are matters of degree of cultural influence and universality or near-universality of experience. Consequently, it is preferable to speak of "culture fair" tests rather than "culture free" tests, in connection with tests which utilize materials that do not handicap or favor any segment of the population for whom the test is intended.

The presence of cultural influence in a test that *appears* to be "culture free" was demonstrated in a study made with several tribes of North American Indians.²⁰ The Goodenough Draw-a-Man Test was used. In a group of Hopi Indians, the mean IQ for boys was 123, while for girls it was 102. Zias also showed appreciable differences in favor of boys, whereas in a group of Navahos, the means of boys and of girls were very nearly equal (107 and 110). The sex differences or similarities within each tribe are attributed to sex differences or simi-

²⁰ R. J. Havighurst et al., "Environment and the Draw-a-Man Test: the Performance of Indian Children," *Journal of Abnormal and Social Psychology*, Vol. 41, 1946, pp. 50-63.

larities in training and experience within each culture. Boys and girls are trained to observe different aspects and details of their environment and are taught different types of drawing. The two sexes have different functions in their group; these functions are reflected in differentiated training; the differences in training are reflected in differences in performance.

Since every person must develop in an environment of some kind, his skills, information, repertory of responses, modes of thinking, etc., are to some extent culturally conditioned. Some psychological tests are more "culture fair" than others. At this point we recall again Binet's principle that a test of intelligence should be consonant with the milieu of those who are to be measured by it.

II.

VERBAL AND MIXED GROUP SCALES OF MENTAL ABILITY

THE scales presented in this chapter are either entirely or predominantly verbal in content. The proportion of nonverbal materials varies in different scales; but it will be noted that even in those instruments that contain appreciable portions, symbolic materials (language and number) predominate.

Since there are many scales that come within this classification, it is neither possible nor necessary to describe and analyze all of them. It is the purpose of this chapter to present a sufficient number of representative scales so that the student can know their characteristics and content, their quality, advantages and disadvantages, their similarities and differences. The details of content are presented in order to provide the student with a clearer conception of the psychological processes being tested than would otherwise be the case. The statistical details for each test will enable the student to see clearly the techniques of standardization used and the degree of success achieved, upon which evaluations of these instruments must rest.

CALIFORNIA TESTS OF MENTAL MATURITY¹

Contents. The revised series provides scales on five levels: preprimary (kindergarten and entering grade 1); primary (grades 1-3); elementary (grades 4-8); intermediate (grades 7-10 and

¹ By E. T. Sullivan, W. W. Clark, and E. W. Tiegs. Published by California Test Bureau, Los Angeles, 1951.

adult); advanced (grade 9 and adult). All of these are designed to test the same "mental factors"; hence, since the series covers the wide age range indicated, it is essential that the content of each scale be adapted to its particular level, as regards difficulty and form. Thus at the earliest levels there must be emphasis upon nonverbal materials, with minimum requirements made upon word knowledge and number concepts. The later levels increase in their demands upon word knowledge, number concepts, and reading of and reasoning with numerical and verbal materials, while increasingly complex nonverbal materials are also retained.

We shall describe only the elementary scale (grades 4-8), since it is representative of the entire series.

The elementary scale consists of twelve subtests, grouped under five headings, or factors, as follows:

Memory—

1. Immediate recall: series of words pronounced in pairs; then only the first word of each pair is repeated, and the subject is to recall the second word of the pair.

2. Delayed recall: a story is read to the subjects. Thirty minutes later they are given a series of multiple choice items to test the extent to which details of the story are recalled.

Spatial relationships—

3. Sensing right and left: 20 pictures of hands and feet in various positions. The task is to discriminate between right and left.

4. Manipulation of areas: spatial patterns of a variety of forms and in different positions to be manipulated, to test spatial imagery.

Logical reasoning—

5. Opposites: 15 sets of drawings, showing five objects in each set. The first is the usual "given" object; the testee selects from the other four the one that is the opposite of the first.

6. Similarities: the well-known classification test, using 15 sets of drawings. The first three in each set are alike in some respect. The task is to select the similar item in the remaining four drawings.

7. Analogies: the familiar test of relationships, using drawings. The first two stand in some relationship (a hat and a man's head); a third item is given (a shoe). The task is to select a fourth item

that bears the same relationship to the third as the first does to the second.

8. Inference: a major and a minor premise are given. The task is to select a logical conclusion from among several alternatives.

Numerical reasoning—

9. Number series: each series of numbers increases or decreases according to a principle. The testee has to discern that principle.

10. Numerical quantity: number concepts using coins. The subject indicates how many coins of each denomination are required to make up a specified sum.

11. Numerical quantity: arithmetical problems.

Verbal concepts—

12. Word similarities: 100 given words. In each item, the task is to select one word, from four, that is synonymous, or nearly so, with the given word.

This scale also provides three optional subtests: of visual acuity, auditory acuity, and motor coordination. These are not included in the scoring. They may be used to learn whether, in respect to these functions, the subject will or will not be handicapped on the subtests that are scored. The use of these three preliminary tests can be very helpful in identifying persons who would be under a disadvantage in taking a group test, and who should, therefore, be examined individually. When an individual scale is used, existing handicaps can be overcome, in part at least, and can be taken into account in the interpretation of the testee's performance and score.

Scoring. An aspect of these scales worthy of special note is the method of scoring. The score is obtained for each of the divisions (e.g., immediate recall, delayed recall) under each of the factors (memory, etc.). For each separate score, a mental-age rank is found and plotted on a scale, or profile. The separate scores of each division are then added to give the score for the particular factor (in this instance, memory), which is also plotted on the profile. The scores of all factors are then added to yield the total score, from which the usual indices may be derived. Also, the appropriate subtest scores are added to obtain a rating on language factors (subtests 2, 8, 11, 12);

then the scores on the remaining subtests are added to get a rating on nonlanguage factors.

Thus these scales enable the examiner to obtain: (1) ratings on the several subtests: (2) separate mental ages and intelligence quotients for verbal subtests combined, nonverbal subtests combined, and total score of all subtests. This type of scoring and profile permits ready analysis of the subject's weaknesses and strengths, consistencies and inconsistencies, in the types of mental operations being tested, assuming, of course, that the test materials are valid and the reliability is high.

Validity and Reliability. Although the norms for this scale are “. . . based on a controlled (stratified) sampling of over 125,000 cases . . .”² reliability statistics were based on only 725 pupils in grades 4 to 6 “. . . in representative school districts.” The split-half method (presumably odd-even) was used, the coefficients having been corrected by the Spearman-Brown formula. The results were:

“total mental factors”: coefficient = .95;
standard error of measurement, 3.5 IQ points.

“language factors”: coefficient = .94;
standard error of measurement, 3.9 IQ points.

“non-language factors”: coefficient = .92;
standard error of measurement, 4.5 IQ points.

.87 (spatial relationships) to .92 (memory).

Standard errors of measurement in the subtests range
from 4.5 to 5.8 months of mental age.

On the whole, these coefficients fall within the range of quite satisfactory correlations, particularly for the two major divisions and for the total score. The standard errors of measurement, in terms of IQ points, also compare favorably with those of the sounder tests. Although the errors of measurement for the subtests are somewhat larger, this is what would be expected. Once again, greater reliability of wide rather than narrow sampling of performance is demonstrated.

It is in respect to the accepted validity criteria that the authors of these California scales do not provide adequate data.³ The authors'

² *Manual*, p. 22. The socio-economic distribution of the sampling is not reported in the manual.

³ In a mimeographed prospectus kindly furnished by the tests' authors, there

purpose, to begin with, is to measure most of the kinds of mental processes sampled by the Binet scales, based upon an analysis of the "conceptual framework" of the Binet; but no correlational data with the Stanford-Binet are given in the manual. Unpublished data, provided by the publishers of these scales, show the following results (Table 40) obtained with the Elementary scale.

TABLE 40

Correlations between California Elementary Scale and S-B

California		Stanford-Binet		N	r
Md. IQ	S.D. IQ	Md. IQ	S.D. IQ		
107.3	28.5	108.2	29.3	283	.94
122.0	27.9	128.5	28.8	182	.93
79.2	22.7	79.5	21.6	101	.90

Several significant facts and inferences emerge from the foregoing table. First, there is close correspondence between the two scales in respect to medians and standard deviations of intelligence quotients. Second, the groups are not representative of the population for which the scale is intended, since the medians are significantly above or below 100, and the S.D.'s are very much larger than would be found in an unselected population (namely, about 16). Third, due to the very wide range of ability, as represented by the very large standard deviations, the probability is that the correlation coefficients are significantly higher than they would be for a population of narrower and typical range. Fourth, it would be desirable to determine this aspect of validity separately for each age group, since close correspondence at one age level or for combined age levels does not necessarily assure equally close agreement at other age levels.

The manual furnishes data on the intercorrelations of the subtests, based on 1048 cases in grades 4 to 6. In view of the authors' previous preference for the group factor theory of intelligence (stated in earlier editions of the scales) and of the generally low intercorrelations of the

are extensive data dealing with the 1947 short form of the advanced scale. This prospectus reports significant correlations between the California and other group scales. For the "language factors" of the California scale, the coefficients are largely in the .70's, whereas for the "non-language factors" they are in the .50's. Correlation coefficients with school grades vary, as they do with all tests of intelligence, and are higher for the "language factors." For these factors the correlations are, for the most part, in the .50's, .60's, and .70's. Similar data are not provided for the elementary scale described herein.

present subtests, we may conclude that these coefficients are regarded as one evidence of the scale's validity: namely, that the subtests have relatively little in common regarding measured factors; hence they satisfy one requirement of the accepted theory.⁴

The subtest intercorrelations may be summarized as follows:

range of subtest intercorrelations: .25 (spatial relationships and numerical reasoning) to .60 (memory and verbal concepts)

seventy percent of the coefficients are below .50; fifty percent are below .40

range of correlations of subtests with language subtests: .35-.95 (in part self-correlations)

range of correlations of subtests with nonlanguage subtests: .55-.78 (in part self-correlations)

range of subtest correlations with total scores: .60-.86 (in part self-correlations)

Regarding validity, one may draw some inferences, however, from data provided in other connections. (1) Median IQ for the entire population sample is 100, with a standard deviation of 16. (2) The population samplings at the higher educational levels show a progressive increase in medians and decrease in standard deviations, until at the college sophomore level these are 114.5 and 13.5, respectively, and at college graduate level they are 124 and 12. Such progressive increases are to be expected, since advancing educational levels are more or less selective as regards intelligence.

Even when a scale is validated within a theoretical conception and framework, it is still necessary to subject the scale to other validating procedures to determine if it actually serves the purposes for which it is intended. In other words, does the scale "work"? What is its predictive efficiency? The necessity of answering these questions is the reason why all tests of intelligence should be validated against accepted criteria of intelligent behavior. Just devising test materials that satisfy a theoretical conception gives no assurance of the scale's practical and predictive validity.

⁴ The reader will recall that the original Binet scales and the Stanford-Binet are based upon a general factor theory.

In the test manual, page 4, it is stated that "The total mental factors score has been found by the authors and other investigators to correlate as high or higher with the individual Stanford-Binet than any other mental ability test." No data relevant to this aspect are given in the manual; but the several correlations reported above were provided in a personal communication.

TERMAN-McNEMAR TEST OF MENTAL ABILITY⁵

Contents. This scale (in two equivalent forms) is intended for use primarily in grades 7 through 12, though norms are provided from the age of 10 years through 19 years, 11 months. The scale consists of seven subtests: information, synonyms, logical selection, classification, analogies, opposites, and best answer.

Items from
Terman-McNemar Test of Mental Ability
(World Book Co. By permission)

Information:

Polo is a kind of

- (1) disease (2) work (3) bear (4) game
(5) language

Synonyms:

- Comic—(1) clumsy (2) laughable
(3) universal (4) tricky
(5) peculiar

Logical selection:

An orchestra always has

- (1) violinists (2) piano (3) musicians
(4) saxophone (5) singers

Classification:

- (1) Catholic (2) Methodist (3) Presbyterian
(4) Republican (5) Baptist

Analogies:

Zoo is to animal as aquarium is to:

- (1) birds (2) fish (3) bees (4) statues
(5) butterflies

Opposites:

- Exit: (1) emit (2) transcend
(3) entrance (4) origin
(5) arrival

Best answer:

The saying, "Idle brains are the devil's workhouse," means

- (1) The devil is lazy.
(2) People who are idle get into trouble.
(3) Many hands make light work.
(4) The devil works with his brains.

⁵ By L. M. Terman and Q. McNemar. Published by World Book Company, 1942. This scale is a revision of the Terman Group Test of Mental Ability which was published in 1920.

The content of the scale is quite homogeneous in that it is entirely verbal in character. The scale is thus consistent with Terman's definition of intelligence: that is, the ability to deal with symbols and abstractions.

The authors subscribe to the general factor (*g*) theory of intelligence. They hold that the general factor is best tested by means of materials using symbols and abstractions. In order to achieve a high degree of homogeneity in test materials, they even omitted from their scale arithmetical and numerical types of subtests, which are widely regarded as being very good tests of the general factor. The authors state the reason for their selection of materials as follows: "More homogeneous material has been used in order to have a test more highly saturated with a common factor or ability. Thus, the exclusion of arithmetical and numerical subtests means that the scores of any two individuals are more nearly comparable qualitatively; i.e., they lie along the same continuum. This continuum may be characterized as general verbal intelligence."⁶ The usefulness of this scale, therefore, is restricted to subjects who are not laboring under a language handicap and to situations wherein "verbal intelligence" is required as the sole or major ability.

Validity and Reliability. Reliability of this scale is presented in terms of three familiar methods: the split-half method (correlation of scores on odd-numbered and even-numbered items), the interform method (correlation of scores on the two forms given to the same subjects), and the probable error of measurement. The split-half reliability coefficient was .96 (279 cases, grades 7 through 9), while the interform reliability coefficient was .95 (239 cases, grades 7 and 9). When an age range of only one year was taken (13-6 to 14-5), both coefficients were .96. The probable error of measurement was found to be 2.2 standard score points. (In terms of the scoring units of the scale, this probable error can be considered small.)

Validity of individual items was determined primarily on the basis of percent of pupils passing each item in the successive grades—in other words, the extent to which each item differentiates between groups at different levels of maturity. A second criterion of item validity was the correlation of each item with total scores.⁷ No item

⁶ *Manual of Directions*, p. 1.

⁷ For this purpose, the "tetrachoric correlation" was used. This technique differs from the more familiar one used in correlating two sets of variable

was retained if it yielded an average coefficient of less than .30; in fact, ninety percent gave correlations of .40 or higher, with an average of .53. Thus, for validity, the final selection of items was made on the basis of item difficulty and degree of relationship of item performance to total scale performance.

Scoring. Two statistical factors are at the base of the rating method used: (1) a single origin; and (2) comparable units in all parts of the scale. The reader is already familiar with the fact that raw scores do not provide comparable units throughout a scale. For the Terman-McNemar scale, this problem is met by a type of standard-score scale which uses the median of the 14-year age group of the national standardization population as the origin, and the standard deviation of this age group, arbitrarily made 16 points, as the unit of measurement. "Scores on this scale for all age groups are thus measured from a single origin and provide comparable units throughout all parts of this scale."⁸

In other words, the authors of the scale have devised a method of scaling which is a variant of the familiar standard score. The 14-year age group is taken as the standard; norms of other age groups are given in terms of standard scores of the 14-year group. Thus, for age 14, the median raw score was 76; this was arbitrarily called a standard score of 100. The raw score standard deviation (actually about 27) was assigned an arbitrary value of 16. No reason is cited for the choice of this particular number. It may be noted, however, that 16 was the most typical value of the standard deviation for the IQ's of the 1937 revision of the Stanford-Binet and thus is a convenient number to use in a group scale which attempts to measure the same type of ability. Having assigned this arbitrary value, the authors then proceeded on the assumption of normal distribution and from a table determined the standard scores that would correspond to various raw scores in a distribution that had a mean of 100 and S.D. of 16. For example, the average raw score of the 13-year group (13 years, 0 months) is 63. In a normal curve with a standard deviation of 16 points and a mean of 100, the score of 63 would have as its equivalent

scores in that the underlying assumptions are different; but interpretation of obtained coefficients is approximately the same. For an explanation of the tetrachoric method, see any standard textbook in statistics.

⁸ *Manual of Directions*, p. 4.

a standard score value of 93, which then becomes the norm for the age 13 years and 0 months.

The procedure for finding an individual's standard-score rating on this scale amounts to this: the raw score is obtained; each raw score is found in a table which gives its standard score equivalent; each standard score is found in another table giving its equivalent mental age. The IQ, however, is not found by means of the usual formula ($IQ = MA/CA$). In fact, the mental age is not used in calculating IQ for this scale. Instead, the authors provide a table for finding what is called the "deviation IQ," so called because "Basically the procedure for computing deviation IQ's requires that the difference be found between the [individual's] obtained standard score and the average standard score for other individuals of the same age. This difference or deviation is then interpreted directly in terms of IQ [from a table]. This can be done because both IQ's and the normalized standard scores are distributed normally."⁹ In other words, Terman and McNemar assume that raw scores and IQ's are both normally distributed. When this assumption is made, it is possible, through knowledge of the characteristics of the two curves, to transmute one set of scores directly into the other.

We have presented this scale's method of deriving IQ's in some detail because it is essential that the student of psychological testing be aware of the several techniques and of the fact that indexes called by the same name are not always derived in the same manner. First, we have the original and most common method of finding IQ: mental age divided by chronological age. Second, there is the method used with the Bellevue scale. And third, there is the type of deviation IQ described above.

TESTS OF PRIMARY MENTAL ABILITIES¹⁰

Contents. The Thorndikes have published a series of group tests variously known as The Chicago Tests of Primary Mental Abilities and The SRA Primary Mental Abilities. The latter are more recent scales, shorter and less satisfactorily standardized and reported than the original Chicago PMA tests, and on the whole inferior to the earlier versions. The following description, therefore, will be limited

⁹ *Manual of Directions*, p. 9.

¹⁰ By L. L. and T. G. Thorndike. Published by Science Research Associates; several forms, 1938-1950.

to the Chicago PMA tests (1943), devised for ages 11–17. This particular scale will serve our purpose, since all of the scales in the Thurstones' series are based upon the same psychological and statistical principles, even though all are not of equal merit.

The PMA scale for ages 11 to 17 is constructed upon the group-factor theory of mental ability; that is, upon the theory that intelligence consists of the operations of certain distinguishable and relatively independent mental functions. (See Chapter 3.) The "primary abilities" to be tested by means of this scale are those which L. L. Thurstone and his collaborators report as having been isolated by factorial analysis. The "primary factors" measured are six: number facility, verbal comprehension, spatial perception, word fluency (extent of word associations as distinguished from verbal comprehension), reasoning, and rote memory. Each of these is measured as indicated below:

Number facility, by tests of addition and multiplication.

Verbal meaning, by one test of vocabulary (the familiar multiple-choice form) and one of supplying words to fit given definitions. In each item of the latter, five letters are provided, one of which is the first letter of the correct word.

Spatial perception, by tests in which designs and geometric figures, differently rotated, are to be identified as being the same as or different from a given design or figure.

Word fluency, by two tests—one requiring that, within a time limit, as many words as possible be written, beginning with a given letter; the other requires that as many four-letter words as possible, beginning with another letter, be written within a time limit.

Reasoning, by two tests involving, in one, perception of the patterns within series of letters of the alphabet, and, in the other, perception of the patterns within letter groupings.

Rote memory, a names test. There are twenty cards, on each of which are a first and last name. The cards are exposed consecutively, each for fifteen seconds. The subjects are then required to pair off each last name with its correct first name (chosen from seven names in multiple-choice form).

Validity and Reliability. By means of the split-half method, reliability coefficients were separately calculated for five of the six subtests for grades 6, 8, 10, and 12, with approximately 200 subjects at each half-grade level. No reliability coefficients are provided for "word fluency," since the scores on this subtest do not lend themselves to the split-half method. Instead, it would be necessary to use the re-test

COMPLETION

Read the definition below. Think of the word which fits the definition. The **first** letter of the word is in the row of letters under the definition.

The first meal of the day.					
A =	B	C =	D =	E =	

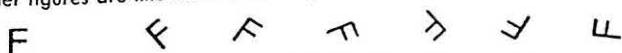
The word is "Breakfast." "B" is marked because it is the first letter of the word "Breakfast."

Do the following example:

A place or building for athletic exercises.					
D =	G =	H =	T =	V =	

FIGURES

Look at the row of figures below. The first figure is like the letter **F** which is right side up. All the other figures are like the first but they have been turned in different directions.



Satisfy yourself that all of these figures look like the first one if they are turned right side up.

Now look at the next row of figures. The first one looks like an **F**. But none of the other figures would look like an **F** even if they were turned right side up. They are all made backward.



Some of the figures in the next row are like the first figure. Some are made backward. The figures like the first figure are marked.

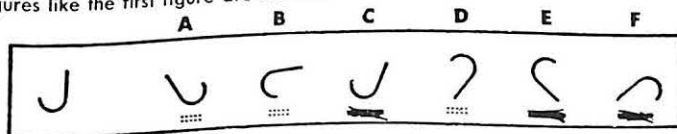


FIG. 11.1. Items from *Chicago Tests of Primary Mental Abilities*. Science Research Associates. (By permission.)

CARDS

Here is a picture of a card. It looks like an L, and it has a hole in one end.



The two cards below are alike. You can slide one around on the page to fit the other exactly.



Now look at the next two cards. They are different. You cannot make them fit exactly by sliding them around on the page.



Here are more cards. Some of the cards are marked. The cards which are like the first card in this row are marked.

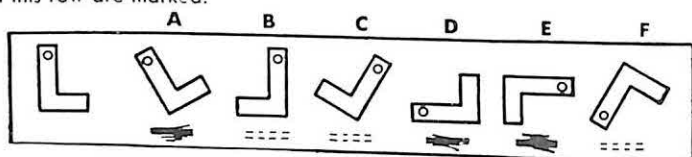


FIG. 11.2. Item from *Chicago Tests of Primary Mental Abilities*. Science Research Associates. (By permission.)

method; but this had not yet been done at the time the scale and its manual were published (1943). With the exception of rote-memory reliabilities, the coefficients reported are all between .95 and .98 (corrected by the Spearman-Brown formula). These are very high. The memory tests, however, showed coefficients of low reliability, ranging from .63 to .82. Seven of these coefficients were in the .60's, two in the .70's, and three in the .80's.

Validity of the Chicago scale is reported in a manner quite different from those which are used with most other scales thus far discussed. Having accepted the group-factor theory, and having analyzed mental abilities into the six "relatively independent" factors already specified, the authors proceeded on the hypothesis that the scale would be valid if the correlations between the primary abilities were relatively low; for such correlations would show relative independence of factors, as required by the theory, and would thus satisfy the underlying group-factor theory of mental abilities. In other words, the *adopted theory* of intelligence is used as the criterion of validity rather than the more usual and generally accepted evidences of mental

development and intelligent activities. Or, to state the same thing otherwise, the criteria of the scale's validity are *internal* rather than *external*. The intercorrelations thus found range from .13 (spatial perception with rote memory) to .58 (reasoning with verbal comprehension). About three fourths of the coefficients are above .30; and about half are above .40.

LETTER GROUPING

Look at the groups of letters below.

AABC

ACAD

ACFH

AACG

Three of the groups have two A's. The group which does not have two A's is marked.

Here is another problem. Three of the groups are alike in some way. Can you find three groups which are alike? Mark the one that is different.

XURM

ABCD

MNOP

EFGH

In three of the groups the letters are arranged in alphabetical order. The first group is not in alphabetical order. You should have marked it to show that it is different.

LETTER SERIES

Study the series of letters below. What letter should come next?

a b a b a b a b

a b c d e f

The next letter in this series should be a. The letter a has been marked in the answer row at the right.

Now study the next series of letters and decide what the next letter should be. Mark the letter in the answer row at the right.

c a d a e a f a

a c d e f g

You should have marked the letter g.

FIG. 11.3. Items from *Chicago Tests of Primary Mental Abilities*. Science Research Associates. (By permission.)

An individual's raw scores obtained on this scale, for each of the "primary abilities," are converted into percentile ranks which are then plotted on a profile. In the first versions of the tests, no mental age, intelligence quotient, or other *general* index was obtained, "Since the principal purpose of the present test is to obtain a profile of the six

primary mental abilities for each child . . .”¹¹ It was the view of the authors of this scale that representation of mental abilities by means of a profile is not only consistent with their theory of mental organization and functioning, but that a profile is most valuable for purposes of interpretation of an individual's performance and for educational and vocational guidance.

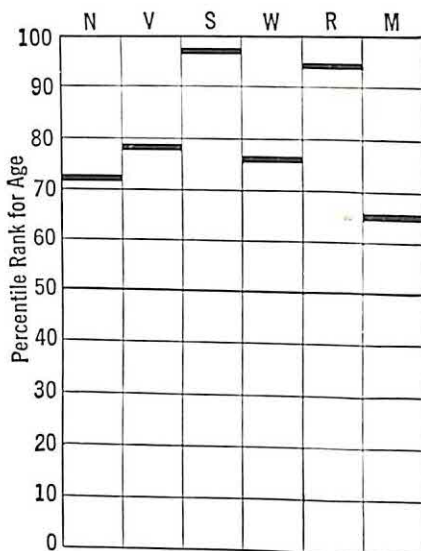


FIG. 11.4. An Individual Profile Chart—Chicago Tests of Primary Mental Abilities. Manual of Instructions, *Chicago Tests of Primary Mental Abilities*. Science Research Associates. (By permission.)

The data thus far available indicate, however, that taken as a whole, as a composite, the tests of “primary mental abilities” are in some instances moderately and in others poorly correlated with scholastic performance at the high-school and college levels. The data show also that the “primary mental abilities,” when considered separately or in “patterns,” do not differentiate significantly between persons interested in or engaged in the study of various professions (engineer-

¹¹ *Manual of Instructions*, p. 19. In later versions of the scales, the authors provide mental age units and IQ equivalents. However, they still regard separate percentile ratings as superior to MA and IQ.

ing, science, linguistic studies, medicine, journalism, etc.).¹² Judgment must be suspended, therefore, regarding the value of these tests in specific rather than general problems in educational and vocational guidance.

KUHLMANN-ANDERSON TESTS (6TH EDITION)¹³

Description. The first edition of this series of tests was published in 1927. Since that time they have undergone extensive experimentation, re-standardization, and improvement so that they are at present among the superior group scales. They are superior in respect to standardization procedure, both intensive and extensive analysis of results, and availability of statistical evidence presented in the manual.

The entire series of scales, of which there are nine, graded according to school level, includes thirty-nine subtests. Each subtest *as a whole* is placed according to its over-all relative difficulty in the age range; and the items *within* each subtest are placed in order of *their* difficulty. Since intelligence levels vary considerably within any single age group, and since there is overlapping among different age groups, there is also overlapping (duplication) of subtests from one scale to the next. Thus, the scales for the several adjacent levels include the subtests as indicated below:

kindergarten, subtests 1-10
grade 1, subtests 4-13
grade 2, subtests 8-17
grade 3, subtests 12-21
grade 4, subtests 15-24
grade 5, subtests 19-28
grade 6, subtests 22-31
grades 7-8, subtests 25-34
grades 9-12, subtests 30-39

The subtest types are not unique, since they include the familiar nonverbal and verbal materials such as sequences, classification, num-

¹² See *Review of Educational Research*, Vol. 14, No. 1, 1944, Chapter 3; Vol. 17, No. 1, 1947, Chapter 2; Vol. 23, No. 1, 1953, Chapter 2.

¹³ By F. Kuhlmann and R. G. Anderson. Published by Personnel Press, Inc., Baltimore, 1952.

ber concepts, form perception and synthesis, word knowledge and facility, scrambled sentences, arithmetical reasoning, verbal analysis, etc. In the earlier levels, the materials are nonverbal; but they develop gradually into scales that are largely verbal and numerical and, finally, that are entirely so.

Thus, to a considerable extent there is, from age to age, overlapping of the functions being measured by this series of subtests. This is quite consistent with the stated purpose of the scale: namely, to measure the levels of general mental development needed to succeed in school work. In this connection, it should be noted also that the authors of the scale do not recommend the use of separate subtest scores as measures of separate psychological functions or for guidance purposes. But, quite appropriately, they do suggest that *significant* inconsistencies among scores on separate subtests furnish evidence of erratic performance, for which causes and explanations should be sought. Implicit in this position of Kuhlmann and Anderson is the acceptance of the general factor theory of intelligence. This acceptance is later made explicit by them in their validation, as will be indicated below.

Validity and Reliability. The usual criteria of validity have been used: test performance of retarded, average, and accelerated pupils; intercorrelations of subtests; correlations of subtests with total scores; means, standard deviations, and ranges of IQ's at the several grade levels; correlations with quality of school work; and power to discriminate among successive age groups (age differences being fractions of a year), upon which heavy emphasis is laid.

These criteria are, for the most part, reasonably well satisfied: the test differences between the known groups of pupils are significant; the mean IQ's approximate what would be expected; for the various age groups, the standard deviations of IQ's range between 9.5 (age 7) and 16.1 (age 14), but more than half of the S.D.'s are between 13 and 16 points; a table of age norms for each of the thirty-nine subtests shows considerable discriminative ability between age increments of less than a year.

Two-thirds of the subtest intercorrelations fall in the .40's and .50's; while of the coefficients for each of the subtests correlated with total score, more than two-thirds fell between .50 and .81. These

coefficients indicate that there is marked communality of functions being measured by most of the subtests but that at the same time, also, some functions beyond the general one are being measured.¹⁴

Reported correlations with school achievement are relatively high, being from approximately .60 to about .80.

Reliability was studied by means of the several familiar methods:

Odd-even reliabilities (uncorrected) varied from .88 to .95.

Standard errors of measurement, after several retestings varied from 5.5 to 6.5 points in IQ.

The test-retest "index of reliability" is .90 for the data from which the standard errors of estimate were obtained.¹⁵

Since there are time limits for these subtests, the appropriateness of the odd-even method of determining reliability might be questioned. The authors have shown, however, that odd-even reliability is just about as high when all time limits are removed as when the tests are timed, the range of coefficients being from .80 to .95. These results indicate that the scales are essentially tests of "power" rather than "speed."

Scoring. The Kuhlmann-Anderson scales use the median mental age method, employed with the Pintner-Paterson Performance Scale (1917). The procedure is this: a mental age is yielded by each of the ten subtests in a given scale; the median of the ten values is taken as the over-all mental age. When this method is used the principle being applied is that this median value is most representative of an individual's general level of performance, especially so since it is not affected by a few extremely high or extremely low subtest scores, if such occur (whereas the *arithmetic mean* is affected by such ex-

¹⁴ Apparently not much significance is attached to correlation with Stanford-Binet results as an indication of validity. No data are presented in the manual on this criterion, although there is brief mention of close correspondence between IQ's obtained with both scales.

¹⁵ The "index of reliability" is to be distinguished from the "coefficient of reliability." The latter is derived from the correlation of two sets of *obtained* scores; while the former is the correlation between one set of obtained scores and a set of *estimated* "true" scores for the same population sample. The "index of reliability" is always the higher of the two, since it indicates the *probable higher limit* of correlation rather than the correlation that has actually been found for the two sets of obtained scores. An "index" of .90 (as reported above) corresponds to a "coefficient" of .81.

tremes). From mental ages, thus found, and chronological ages, intelligence quotients are determined in the usual manner.¹⁶

GROUP SCALES FOR COLLEGE FRESHMEN

Some tests of intelligence have been constructed for the specific purpose of appraising abilities of individuals with special reference to the intellectual demands of college curricula, generally of the kind found in colleges of liberal arts and similar institutions, such as teachers colleges. We shall briefly describe several of these, principally to familiarize the reader with the types of materials included. These tests do not present any new or unusual principles in respect to construction, organization, or interpretation.

American Council on Education: Psychological Examination for College Freshmen.¹⁷ This scale includes the following familiar subtests: arithmetic problems, word definition, figure analogies, same-opposite (word meaning), number series, and verbal analogies. These six tests are grouped into two classes: (1) linguistic tests (same-opposite, word definition, and verbal analogies) which yield an L-score; (2) quantitative tests (the remaining three) which yield a Q-score.

The authors state that these two separate scores may be used in educational counseling; for the linguistic tests have been found to have higher correlations with scholarship in colleges of liberal arts than do the quantitative scores. This result is due to the fact that a very large portion of courses in these colleges make their principal demands upon linguistic activities and thinking. On the other hand, for scientific and technical curricula, the authors hold, the quantitative tests may be more significant.

The raw scores are converted into separate percentile ranks for total score, L-score, and Q-score. Separate distribution tables are provided for men and women. Also, it is not uncommon for a college to prepare its own frequency tables and to calculate percentile ranks of its students on the basis of their performance alone, rather than on the basis of national norms.

¹⁶ The Kuhlmann-Finch Tests (1952) are modeled after the Kuhlmann-Anderson Tests, in respect to principles and content. Published by Educational Test Bureau.

¹⁷ Developed by L. L. Thurstone and T. G. Thurstone. Published by the Educational Testing Service; annual editions. A separate edition for high-school students is also available.

ITEMS FROM
AMERICAN COUNCIL ON EDUCATION, PSYCHOLOGICAL
EXAMINATION FOR COLLEGE FRESHMEN (By permission)

Figure Analogies:

In Sample 3 below, the rule has two parts: "Make Figure B of the opposite color and larger than Figure A." Apply the rule to Figure C and blacken the space which corresponds to the correct answer.

	A	B	C	1	2	3	4	5
3								

You should have blackened the space numbered 1, which corresponds to the large white square.

Verbal Analogies:

In each row of words, the first two words form a pair. The third word can be combined with another word to form a similar pair. Select the word which completes the second pair. On the answer sheet, blacken the space which corresponds to the word you select.

sky-blue	grass-	(1) ground	(2) sod	(3) path	(4) blue	(5) green
ice-solid	water-	(1) hard	(2) fire	(3) iron	(4) liquid	(5) boat

Number Series:

Find the rule in the series below, and blacken the space on the answer sheet which corresponds to the next number.

10	8	11	9	12	10		9	10	11	12	13
							(a)	(b)	(c)	(d)	(e)

The series above goes by alternate steps of subtracting 2 and adding 3. You should have blackened space (e), which corresponds to 13, the next number.

Ohio State University Psychological Test.¹⁸ This scale includes three subtests: same-opposite (word-meaning), word analogy, and paragraph meaning (answering questions on each of a number of paragraphs, to test the subject's comprehension and interpretation of materials read). The scale is clearly and specifically devised to test verbal intelligence.

¹⁸ Prepared under the direction of Herbert A. Toops. Published by Ohio College Association Committee on Intelligence Tests for Entrance, Ohio State University; Form 24, 1950, is most recent. This scale has been used with high-school students also, for whom separate norms have been prepared.

Norms of percentile ranks are provided, based upon scores of freshmen in a large number of Ohio colleges. Separate tables of norms for total scores are given for men and women. Separate norms are also provided for the test of paragraph meaning. The reason, presumably, for the separate norms for paragraph meaning is that thereby a subject's level of performance in respect only to word knowledge may be distinguished from his ability to interpret and think in terms of linguistic symbols.

As in the case of the American Council scale, it is not an uncommon practice for colleges to prepare their own separate norms and percentile ranks when using the Ohio State scale.

College Entrance Examination Board: Scholastic Aptitude Test.¹⁹

This scale has two sections: a verbal and a quantitative (mathematical). The former section includes tests of word-opposites, word analogies, paragraph meaning, and sentence completion (rather long and complex ones); the latter section includes mathematical problems, involving arithmetic, algebra, and plane geometry.

The purpose of these tests is not to examine an individual's knowledge or mastery of subject-matter covered in high school, even though the test items do utilize the tools of learning and thinking acquired there. The verbal section is designed to measure understanding of words, skill in dealing with word and thought relationships, and ability to read with understanding and discrimination. The mathematics section is designed to measure ability to handle quantitative concepts, rather than achievement in the field of mathematics. On the whole, scores on the Scholastic Aptitude Test are considered to be indexes of probable success in subsequent academic work in college courses involving verbal and quantitative materials. The scores on the verbal sections, obviously, are found to have greater predictive value regarding performance in linguistic studies of all types and in social studies; while the mathematics section has greater predictive value in the study of physical sciences and engineering. It appears that while small differences between scores on the two sections, in the case of a given person, have little significance, large differences are of im-

¹⁹ Prepared by the staff of the College Entrance Examination Board; annual editions; not available for use by others. Although this is called a scholastic aptitude test, it has much the same content as other tests of intelligence at this level and is intended to serve the same purpose.

portance in counseling students in regard to their selection of college curricula.

Raw scores on this scale may be converted into percentile ranks, according to national norms. Or, again, each individual institution may set up its own frequency distribution of scores and corresponding percentile ranks.

ITEMS FROM COLLEGE ENTRANCE EXAMINATION BOARD
SCHOLASTIC APTITUDE TEST (By permission)

Sentence completion:

One of the most prevalent erroneous contentions is that Argentina is a country of agricultural resources, and needs only the arrival of ambitious settlers.

- | | |
|-----------------|-----------------|
| (1) modernized | (2) flourishing |
| (3) undeveloped | (4) waning |
| (5) limited | |

Precision of wording is necessary in good writing; by choosing words that exactly convey the desired meaning, one can avoid

.

- | | |
|------------------|-----------------|
| (1) duplicity | (2) incongruity |
| (3) complexity | (4) ambiguity |
| (5) implications | |

Yale Educational Aptitude Battery.²⁰ This series of tests is made up of seven parts: (1) verbal facility (the verbal section of the College Board Scholastic Aptitude Test); (2) linguistic aptitude (as measured by a test of artificial language); (3) verbal reasoning (logical inference and deductive judgment); (4) quantitative reasoning ("ability in manipulating hypothetical quantitative data so as to perceive relations or principles characterizing them and derive 'laws' analogous to, yet different from, those actually encountered in the study of the natural sciences"); (5) mathematical aptitude (mathematical problems similar to those in the College Board test); (6) spatial visualizing

²⁰ A. B. Crawford and P. S. Burnham, *Forecasting College Achievement* (New Haven: Yale University Press, 1946). The reader will note that the authors of these tests have chosen to call them an *educational aptitude* battery. The contents of these scales, however, are very much the same as or identical with the contents of other scales variously designated as tests of intelligence or of mental abilities, or as psychological examinations. Because the Yale battery is similar to these others not only in content but in purpose as well, it is included here rather than in the chapter on aptitude tests.

("representation of three-dimensional forms by two-dimensional figures through projections and block-counting"); (7) mechanical ingenuity ("problems in gear or pulley movements, structural stability, and mechanical operations").

The reader is already familiar with these types of items from samples of other scales, except numbers 2, 4, and 7. The last of these is generally found in scales designed to test only specialized "mechanical aptitudes," so it will not be illustrated at this point. (See Chapter 12.) The following items illustrate tests 2 and 4, respectively.

ARTIFICIAL LANGUAGE

Vocabulary:

I—vlu
 he, it (nom.)—wes
 to be—jahviz
 to read—skraliz
 to have—dromiz
 good—zeyt
 book—stetsleit
 word—gleit

Rules:

1. Articles are not used in the artificial language.
2. Verbs are not conjugated for person and number.
3. Future—prefix *bli* to the verb.

Sample (to be translated):

	A	B	C		Answers
	I	have	a book		
A. (1) wes, (2) polvlu, (3) vlu, (4) polwes, (5) vlul				1	2 3 4 5
B. (1) dromiz, (2) jahviz, (3) amdiz, (4) somiz, (5) binotiz				1	2 3 4 5
C. (1) gleit, (2) zepoldeit, (3) zeyt, (4) stetsleit, (5) oveit				1	2 3 4 5

Quantitative Reasoning

A	B
8	2
32	4
18	3
200	10
50	5

The problem is first to discover and state the relationship between the paired numbers. In this illustration it is: $A = 2(B^2)$. Then: when $A = 72$, what is the value of B ?

The Yale battery differs from the other scales for college freshmen principally in the following respects: it contains a more diversified and a larger number of types of items; it attempts to predict college success in more specialized areas of study (verbal, scientific, engineering); and it provides the tables, in terms of points and percentile-rank equivalents, for obtaining a profile graph for each person tested. On the whole, this battery is intended to reveal individual differences in major areas and in educational promise at the higher levels which, most persons would grant, depends to a very large degree upon those mental functions that have been designated as constituting intelligence.

Evaluation of Group Scales for College Freshmen. The reliability coefficients of the four scales discussed above are of the general order of .90. The method usually employed is the familiar one of correlating scores on odd-numbered items with those on the even-numbered ones.

The intercorrelations presented in Table 41 are representative of the subtests used at the level of college freshmen.

Since scales for college freshmen are designed for a specific purpose—selection of promising students and prediction of college achievement—their validity depends upon the success with which they perform this task. The scores on these scales, therefore, have been correlated with marks obtained in college courses.²² The medians of the coefficients found in recent years are generally in the neighborhood of .50 and .60.

Test scores have been used to predict survival in college, disregarding actual marks. It was found that the scales are useful in helping to identify individuals at the two extremes: namely, those who are most likely to complete their undergraduate work and those who are least likely to do so. The test scores do not differentiate very well between individuals in the middle range of the distribution (approximately the mid-sixty percent), so far as scholastic achievement is concerned. The reasons for this are that this middle group is fairly homogeneous, that numerous nonintellectual elements affect students' course marks, and that college marks themselves are neither objective nor entirely reliable. The results of correlational and the "survival" studies are not such as to warrant exclusive use of psychological tests in selection and elimination of candidates for admission to college.

²² See *Review of Educational Research*, Vol. 14, No. 1, 1944, Chapter III; and the first numbers in 1950 and 1953; also Crawford and Burnham, *passim*.

TABLE 41

ACE Psychological Examination:
College Edition, 1948²¹
(385 Cases at One College)

Intercorrelations, Means, and Standard Deviations of the Six
Subtests, Q, L, and Total Score

	1	2	3	4	5	6	7	8	9
(1) Arithmetic	.432	.558	.427	.297	.450	.739	.439	.643	
(2) Figure Analogies .432		.488	.349	.285	.529	.836	.442	.690	
(3) Number Series . .558 .488			.323	.283	.498	.847	.421	.682	
(4) Completion 427 .349 .323				.658	.574	.436	.829	.757	
(5) Same-Opposite . .297 .285 .283 .658					.563	.351	.904	.766	
(6) Verbal Analogies .450 .529 .498 .574 .563						.611	.822	.836	
(7) Q-Score 739 .836 .847 .436 .351 .611							.530	.827	
(8) L-Score 439 .442 .421 .829 .904 .822 .530								.915	
(9) Total Score 643 .690 .682 .757 .766 .836 .827 .915									
Mean	8.16	15.96	16.58	16.53	21.16	26.86	40.70	64.56	105.25
σ	2.91	5.34	4.85	4.14	7.16	5.62	10.72	14.94	22.53
Percentile rank of means							44	50	47
Results for all colleges as reported in Norms Bulletin for 1948, College Edition									
Mean						41.56	64.38	105.91	
σ						11.39	16.12	24.65	

But they do make a valuable contribution when used in conjunction with other information and criteria concerning each candidate, such as high-school marks, performance on subject-matter examinations, and ratings by teachers.

ARMY GENERAL CLASSIFICATION TEST²³

This scale, in various earlier versions, was used in the Army in World War II to classify men and women according to their abilities to learn military duties. Since much of the learning in modern military service is of a technical kind, emphasis was placed upon verbal comprehension, quantitative reasoning, and spatial perception. The AGCT, therefore, adapted and utilized conceptions of testing and

²¹ Communication from Educational Testing Service. (By permission.)

²³ Published by Science Research Associates, Chicago, 1947. See Staff, Personnel Research Section, Classification and Replacement Branch, Adjutant General's Office, *Personnel Classification Tests*, War Dept. Technical Manual, TM 12-260 rev., Washington: U. S. Govt. Printing Office, 1946.

types of materials that had been developed and in use for a long time prior to the war. Three types of test items were employed to measure the three processes, respectively: vocabulary, arithmetical problems, and block counting. The items are presented in spiral form: a group of vocabulary items, then a group of arithmetical problems, then blocks. The sequence is repeated a number of times, each group of items being of greater difficulty than the preceding one. They are of the usual multiple-choice type.

Scoring. The raw scores might range from zero to 150. These are converted into a standard score, so arranged that the mean is 100 and the standard deviation is 20. Tables are also provided for conversion of raw scores into percentile scores.

TABLE 42
Distribution of AGCT Standard Scores ²⁴
(N = 160,000)

Score	Percent of Sample
41-59	4
60-89	23
90-109	31
110-129	33
130-161	9

Validity and Reliability. Using the odd-even method, the mean of the reliability coefficients (corrected) was approximately .95. (N varied from 639 to 3856.) When the test-retest method was used, the reported reliability was .82.

The AGCT scores correlated .73 with number of school grades completed. It correlated as follows with other tests: Army Alpha, .90; Otis Higher Mental Ability Examination, .83; American Council on Education Psychological Examination, .79. That the AGCT scores are not in part a function of chronological age is shown by a correlation of .02 with CA (N = 4330).

Test scores were correlated with training-school marks in several hundred training programs. The coefficients that follow are based upon groups of trainees who had been preselected for a particular

²⁴ From *Examiners Manual*, AGCT, p. 3. Science Research Associates. (By permission.)

school on the basis of education and civilian occupation. The coefficients are lower, therefore, than they would have been if an entirely unselected group of trainees had been used. Correlations were: clerical trainees, .40; airplane mechanic trainees, .35; sheet metal trainees, .27; radio operator-mechanic trainees, .32; officer candidates (various services), .40.

The ranges of scores, classified according to civilian occupation of the subjects, show such extensive overlapping of groups (interquartile range, also 10th and 90th percentile range) that occupational differentiation would be extremely doubtful, except in instances where occupations being compared are very widely separated on the scale.

The army versions of this test were constructed for the purpose of rapid and rather crude screening of huge numbers of inductees, under pressure of an emergency situation. The tests did this job satisfactorily. But adaptation of these tests to civilian purposes is a very doubtful procedure; for superior single scales and batteries of tests are available. These scales and batteries, which often require much time and careful interpretation, are to be preferred in civilian situations that are free from pressures for speed and in which extravagance with human and material resources is discouraged.

MILLER ANALOGIES TEST ²⁵

Originally devised in 1926, this test has been developed to measure scholastic aptitude at the graduate-school level. It consists of one hundred items, with a time limit of fifty minutes; but the speed factor is said to be of negligible importance. The test includes analogies covering a wide variety of fields of learning and specialization. Although some quantitative as well as verbal materials are used, the items are predominantly verbal in character. Also, the relationships to be discerned within each of the items are not uniform; thus the subject must reorient his analysis with each item. (It is not possible to illustrate or describe the items, since the test is not circulated and its use is controlled.)

Items were retained for their differentiating ability as determined by an item analysis of results obtained with 1241 college seniors.

²⁵ By W. S. Miller. Published by The Psychological Corporation, 1947. A newer form (H), 1950, is also available. The 1947 and 1950 forms are practically equivalent.

These items were then administered to 770 entering graduate students at the University of Minnesota. The latest revision was made on the basis of an item analysis of their scores, the items being arranged in their order of difficulty as determined by percentage of errors. These students represented a wide range of major fields of study.

Validity and Reliability. Coefficients of reliability, found by the odd-even score method, for three groups of graduate students, were .93, .93, and .92, as corrected by the Spearman-Brown formula. The numbers were 100, 162, and 125.

Validity is determined largely by ability of the test to predict success in graduate study. The correlation coefficients between the Miller Analogies Test and numerical marks, as reported in the manual, are restricted to University of Minnesota graduate students in one field: namely, education. For course grades, the twenty-four coefficients ranged from .14 to .78. (With one exception, all were .35 or above.) The median was .54. Ten correlations with grades in final comprehensive examinations ranged from .28 to .54, with a median of .39. Although correlation coefficients are not reported, the data show an increase in mean test-score as honor point-ratios increase. Validation studies have been made subsequently at other universities. The correlations found at these institutions, in other fields of graduate study, were of approximately the same magnitude.

When the analogies-test scores were correlated with average scores on seven parts of the Graduate Record Examination (mathematics, physics, chemistry, biology, history, government and economics, literature and fine arts) the coefficients were much higher. These ranged from .64 to .84, with a median of .775. Correlations with the advanced Graduate Record Examination (fields of specialization) yielded coefficients of .81 and .79 for major students, respectively, in chemistry and languages-and-literature. Correlations with the verbal parts of the Graduate Record Examination were from .74 to .81, the median being .80.

This array of coefficients adds up to the conclusion that the Miller Analogies Test is a useful additional source of evidence in regard to a person's ability to pursue graduate study. Combined with results of the Graduate Record Examinations, the value of each is increased. The lower coefficients obtained with actual course grades are very

probably due to the variations in marking, the differences in difficulty of courses, specialization of interests and abilities and influence of nonintellective factors that affect a student's level of performance.

This analogies test is a carefully constructed instrument, devised for use with a difficult problem: namely, differentiation among individuals in a rather highly selected group. The ability tested by means of these analogies is one factor in the prediction of success in graduate studies: namely, ability to learn verbal and other abstract concepts and course materials. What the analogies do not evaluate (nor do they purport to do so) is original thinking and constructive research ability—both of which should rank very high in graduate studies.

OTHER GROUP SCALES

It is our purpose, in this section, to refer briefly to several scales which will help to give the reader a more nearly adequate conception of representative group instruments.

Institute of Educational Research Intelligence Scale CAVD. This test was developed at Columbia University under the direction of E. L. Thorndike and frequently is designated by his name.²⁶ The scale gets its name (CAVD) from the fact that it proposes to measure intellect specifically by means of four kinds of mental tasks: completion (C), ability to supply words so as to make a statement "true and sensible"; arithmetical problems (A); vocabulary (V), ability to understand single words; directions (D), ability to understand connected discourse as in oral directions or paragraph reading.

The distinguishing feature of this scale is that the items are arranged in order of difficulty, providing seventeen different levels, in each of which the tasks of any one subtest (e.g., C) are of nearly equal difficulty. Also, the steps between levels are of approximately equal difficulty. The lowest level is suitable for three-year-old children while the highest levels are intended for superior adults. Thus the scale is designed to test the same functions in a continuum from early childhood through adulthood.

The range of difficulty in the same category of mental tasks may be illustrated by the following examples.

²⁶ Published by Bureau of Publications, Teachers College, Columbia University, 1925. Norms are not available for levels below ninth grade population.

Completion, lowest level:

You are sitting on a

Completion, one of the highest levels:

Throughout the river plains of northern India, two harvests, and, ... some provinces, ... are each
.....

Arithmetic, lowest level:

Counts two pennies

Arithmetic, one of the highest levels:

A factory earns \$70 a day for its owner when it is working full capacity and \$15 a day when it is working half capacity. In how many days will it earn \$1,000 if two days out of every three are only half capacity?

Vocabulary, lowest level:

"Show me the horse" (to be indicated in a series of pictures).

Vocabulary, one of the highest levels:

Accolade [means] (1) salutation, (2) anchovy, (3) procession, (4) bivouac, (5) acolyte.

Directions, lowest level:

"Make a ring like this," showing act.

Directions, one of highest levels:

A rather long paragraph, entitled "The American State," is read. The following is one of the questions asked: "To what may we attribute the similarity between the plans of certain cities and the arrangement of the States?"

Results obtained with the CAVD scale have been correlated with those obtained by the same persons on other group scales and on the Stanford-Binet. The coefficients obtained were high, and in some instances very high. However, the time required to take the CAVD scale is much longer than that required for the others with which it correlates so highly. Due to the time factor, this scale is not so widely employed as some of the others.

The Thorndike Intelligence Examination for High-School Graduates.

This test is extremely long (requiring about three hours) and laborious to score, as well as expensive.²⁷ It includes tests of highly specialized information, arithmetic, algebra, and paragraph meaning (which constitutes more than two-thirds of the examination). The information and algebra tests may be criticized as being dependent upon highly specific school learning; and the tests of paragraph inter-

²⁷ By E. L. Thorndike. Published by Bureau of Publications, Teachers College, Columbia University; most recent series, 1931-1936.

pretation may be criticized as being dependent upon a high-level vocabulary which requires unusual educational opportunities, either formal or informal. Thorndike, however, intends that the examination shall be used with "candidates who have had good educational advantages and who know English as a mother tongue." The presumption, then, is that, having had these advantages, prospective college students will manifest intellectual competence and promise in terms of their abilities to deal with such mental tasks as are presented in this examination.

Henmon-Nelson Tests of Mental Ability. Though one of the older instruments, this test is still rather widely used.²⁸ Standardized at three educational and age levels (grades 3-8, 7-12, and 12-16), the items are arranged in "spiral omnibus" form. These scales, intended to measure general scholastic intelligence, are weighted with verbal materials. Of secondary importance (in terms of numbers) are the quantitative items; and of distinctly minor significance are the non-verbal (spatial relations) items, there being, for example, only 10 of these in 90 at the level of grades 7-12. The verbal and numerical items are of the familiar variety: synonyms, definitions, analogies, scrambled sentences, number sequences, arithmetical problems, etc.

The Henmon-Nelson scales rank among the superior group tests of general ability, when used with subjects who are not under language or educational handicaps. Although these are scales upon which a great deal of careful work has been done, they are now in need of revision in respect to norms and content.

The following four items illustrate how a "spiral omnibus" test is arranged.

1. Which word does not belong with the others?
(1) Ida, (2) Paul, (3) Lucy, (4) Janet, (5) Edith
2. Better is to *good* as *worse* is to:
(1) very good, (2) medium, (3) bad, (4) much worse, (5) best
3. 1, 6, 11, 16, . . . , . . . , 31. What two numbers should be on the dotted line?
(1) 21 and 26; (2) 17 and 25; (3) 26 and 29; (4) 22 and 27;
(5) 20 and 25.
4. It was raining too hard to out. A word for the blank is:
(1) comment, (2) gather, (3) venture, (4) summon, (5) render.

²⁸ By V. A. C. Henmon and M. J. Nelson. Published by Houghton Mifflin, 1931-1950.

The Pintner Tests. Another series is that of which Pintner was the senior author.²⁹ The *Pintner-Cunningham Primary Test* (1946) covers the range from kindergarten through the first half of grade 2. The *Pintner-Durost Elementary Test* (1940) is devised for the last half of grade 2 through the first half of grade 4. The *Pintner Intermediate Test* (1938) is for the latter half of grade 4 through the first half of grade 9. The *Pintner Advanced Test* (1938-1939) begins with the ninth grade and continues through adult levels.

The Otis Group Intelligence Scale (1919). This test provides two examinations.³⁰ The Primary Examination is designed for the range extending from kindergarten through grade 4. The Advanced Examination extends from the level of grade 5 through grade 12. Otis has also devised several other forms of his tests, which differ from the foregoing chiefly in respect to their mechanical features; some are designed to be "self-administering," and others to facilitate "quick scoring."

There are other group scales which are as well constructed, as valuable, and as widely used as some described in this chapter. It has not been our purpose to present a complete listing and description of group scales; it is unnecessary to do so, since those that have been described herein contain all the essential features of current group tests. It has been our purpose, rather, to familiarize the reader with the organization and content of group scales, their statistical merit, their scoring techniques, and with the major similarities and differences that exist among them.

EVALUATION OF GROUP SCALES

Comparison with Individual Scales. Group scales were developed to permit the testing of large numbers of persons at one time. On the whole, therefore, they are not so useful as are individual scales (e.g., Stanford-Binet and Bellevue) in studying an individual case. For when a group scale is used, it is not possible to observe a person's approach to the solution of problems, nor his behavior under success and failure. Nor is it possible to evaluate the *qualitative* characteristics of his responses, since group scales are scored quite rigidly. Further-

²⁹ Published by World Book Co.

³⁰ Published by World Book Co.

more, it is difficult—in fact, practically impossible—to know whether an individual is exerting his maximum effort when taking a group examination. Thus when a group scale is given, it is possible to report the test results only in terms of numerical indexes (plus profiles, at times), whereas during an individual examination the psychologist is able to make behavioral and qualitative observations of considerable value.

Practically all group scales, below college level, have been validated against individual scales—especially the Stanford-Binet—as one of the principal criteria. This fact in itself is a recognition of the merit of the individual scale, the quality of which the group scale is trying to approach as closely as possible. Other criteria of validity are the familiar ones discussed in earlier chapters.

In discussing the definitions and analyses of intelligence, we stated that one deficiency of all tests is that they do not measure the creative aspects of intelligence; nor do they directly measure the insights that come from experience (“wisdom,” “judgment”), or productive thinking, or the intellectual originality of an individual. This deficiency is more marked in group than in individual scales because of the rigidity of scoring the former.

Theoretical and Statistical Bases. Most group scales are based, implicitly at least, upon the “general factor” theory of intelligence; for most of them undertake to sample a person’s mental activities by means of several kinds of tasks and then to rate the individual by means of a single index. A few scales are based upon the group-factor theory.

Since many group tests, of varying quality, have been published, it is essential that prospective users examine the manuals closely to determine which of these satisfy the standards that should be demanded of them. The reader is already familiar with the standards and methods of establishing reliability and validity. These should be rigorously applied to group tests. In this connection it is essential that the manual state which method was used in determining reliability, especially if the speed factor seems to be a significant one.

Since group tests for children and adolescents are used primarily to assist in dealing with educational problems, it is essential that the scale’s predictive efficiency, with regard to school work and progress, be reported as one criterion of validity.

Scores from a scale as a whole are more reliable and more valid than subtest scores. A distinction should be made, therefore, between subtest reliability and validity, on the one hand, and total scale reliability and validity, on the other. This distinction is especially pertinent when a scale's subtest scores are to be used for differentiating and diagnostic purposes.

The manual should give not only the size of the standardization population sample, but the characteristics of that sample should be specified: namely, geographic and socio-economic distributions, range of ages, range of ability levels, range of school levels, and sex distribution.

Criteria of Evaluation. In evaluating a group scale with a view to its possible usefulness in a given practical situation or in the solution of a theoretical problem, it is customary to use the following criteria:

It must be sufficiently *valid* and *reliable*.

The range of *norms* must be adequate for the group for which the scale is devised.

The *item-difficulty* in each subtest must be of sufficient range to differentiate between the various levels of ability. Individuals at the lowest and highest levels should be able to obtain representative scores.

In general, the *range of ability* to be tested (ages and school grades) should be restricted rather than all-inclusive. By restricting the range, a given number of items and a given length of time can be used for a more thorough and accurate examination than if a scale of the same length were employed to cover a wider range. In the latter instance, the test items would have to be spread more thinly.

Length of the scale must be adequate. In time required, scales vary from about one-half hour to three hours, depending upon levels for which they are intended. The great majority of scales require one and one-half hours or less. Increase in length, to an optimal point, adds to the validity and reliability of a scale; for errors of measurement are decreased (better sampling) as length is increased to an optimal point. Judging from current practices, based upon experiment, optimal lengths appear to be about a half-hour at the level of kindergarten and primary grades, about forty-five minutes at the level of elementary grades, and up to about an hour or an hour and a half at higher levels.

Simplicity of responses is frequently regarded as an asset in group tests. For some purposes—when group trends are sought, rather than individual performance—this is an asset simply because scoring is facilitated. But, as already pointed out, such simplicity and conse-

quent rigidity may limit the value of tests when evaluation of an individual's responses is desired.

Simplicity of scoring is also frequently considered to be an asset, since it is actually a result of simplicity of responses. The same comments apply here as above.

Ease of administering a group scale is desirable. Frequently, group scales have to be given by relatively inexperienced persons; it should, therefore, be possible to train them in a brief time to administer the scale accurately and with precision. Also, simplicity of instructions and procedures in giving an examination to a group reduces the possibility of confusion and misunderstanding on the part of individuals in the group.

The *examiner's manual* should be clear and complete in respect to standardization procedures and results, nature of the content, directions for administering and scoring, norms, and interpretation of results.

The *content* of the tests should be *interesting* to the groups for whom the scale is intended.

The *content* of the tests should be *appropriate* to the subjects being examined. That is to say, the psychologist must determine whether or not, in a given instance, it is desirable to use a scale which is entirely verbal, or entirely nonverbal, or verbal and quantitative, or mixed. His choice of scale will depend upon who are the subjects to be tested and the purpose for which the test is being given.

USES OF GROUP SCALES

Without going into details, we wish at this point only to mention the uses to which group scales have been put.

In schools, they have been used for purposes of general survey, ability classification of pupils, and guidance. Under general survey, studies have been made of the following: range and distribution of mental ability; age and grade overlapping of ability; differences between pupils in various schools within the same community; differences between pupils in different school systems; differences between pupils in the several high-school curricula; the effects of different methods of instruction upon pupils at the several levels of ability; relations between intelligence-test ratings and school achievement in general, and in specific school subjects; and comparisons of city, town, and country children.

In classifying pupils according to ability level for the purpose of differentiated instruction, a test of mental ability is, of course, basic, though it should not be the only criterion.

Since relatively very few schools include a qualified psychological examiner on their staffs, and since extensive individual examination is costly and time-consuming, group scales are being used for most guidance purposes. However, in view of the fact that group-test ratings may indicate only the approximate level of an individual's mental ability, they must be used in conjunction with other available evidence obtained from school records, teachers' reports, objective achievement tests, and interviews. But there is no doubt that psychological test ratings, correctly obtained and interpreted, tell us much more about a pupil's mental alertness and organization of abilities than could be ascertained without their use.

Group tests have been applied extensively to a large number of theoretical and practical problems of psychological, educational, and sociological significance, such as: individual differences in relation to sex, racial, and national membership; mental levels and characteristics of special groups, such as the mentally deficient, the gifted, and the delinquent; employee selection for jobs requiring different levels of ability; family similarities and the inheritance of intelligence; effects of changed environment upon mental level; the nature and course of mental development; the nature and organization of intelligence; constancy of the IQ and prediction of ability; and problems of theory and technique, such as the relationship between "speed" and "power" as aspects of intelligence. Then, of course, there was the vast use of group scales in the armed forces, during World War II, for "screening" and classification of enlisted and commissioned personnel.

The foregoing enumeration is not exhaustive; but it suffices to show the wide range of application of group tests of mental ability; and it explains why tests are under continual scrutiny in an effort to increase their validity and reliability.

APTITUDE TESTS: MECHANICAL AND CLERICAL

DEFINITION AND EXPLANATION

An *aptitude* is a condition or combination of characteristics indicative of an individual's ability to acquire with training some specific knowledge, skill, or set of responses, such as the ability to speak a language, to become a musician, to do mechanical work, etc. An aptitude test, therefore, is a device designed to indicate a person's *potential ability for performance of a certain type of activity of a specialized kind and within a restricted range.*

Aptitude tests are to be distinguished from those of general ability, such as we presented in earlier chapters, and also from tests of skill or proficiency acquired *after* training or experience. They should be distinguished, too, from educational achievement tests which are designed to measure an individual's quantity and quality of learning in a specified subject of study after a period of instruction.

The reader should note that aptitude is differentiated from skill and proficiency. *Skill* means the ability to perform a given act with ease and precision. *Proficiency* has much the same meaning, except that it is more comprehensive; for it includes not only skills in certain types of motor and manual activities, but also in other types of activities as shown by the extent of one's competence in language, bookkeeping, history, economics, mathematics, human relations, etc. We may speak of one's degree of proficiency in any type of performance. On the other hand, when we speak of an individual's aptitude

for a given type of activity, we mean the capacity to *acquire* proficiency under appropriate conditions; that is, his potentialities at present, as revealed by his performance on selected tests which have predictive value.

Furthermore, when we speak of a person's aptitude for a specified activity, we do not make any assumptions regarding the degree to which they depend upon innateness or acquisition. In giving an aptitude test to a person we desire to obtain a measure of his promise or essential teachability in a given area. While we make no assumptions regarding the roles of nature and nurture in this matter, we, as clinicians or guidance counselors, cannot ignore that person's past experience in evaluating his performance on aptitude tests. For example, one method of measuring mechanical aptitude is by means of a mechanical assembly test, utilizing various common objects such as a bicycle bell and a door lock. It is inconceivable that a boy who in the past has had opportunity to manipulate such objects will not achieve a higher score than if he had not had such experience. Testing instruments measuring engineering aptitude include, for example, tests of simple mathematical relationships, scientific vocabulary, common scientific principles, and problems of practical mechanical insight. Here again, an individual's performance will be influenced by his previous experience. This aspect of aptitude testing and interpretation will become clearer as the reader becomes acquainted with the nature and content of aptitude tests.

The principles underlying aptitude tests are the same as those employed with tests of intelligence in respect to sampling of performance, population samples, and standardization techniques (including reliability and validity). Therefore, we shall not present the several aptitude tests in statistical detail. It will be our purpose, rather, to describe the kinds of activities or functions most commonly examined by available tests of this type.

TESTS OF VISION AND HEARING

Quite aside from the general desirability of good vision and hearing, there are numerous occupations and forms of learning in which one or both are essential at a high level; thus, they are aspects of certain aptitudes. Sensory deficiencies, furthermore, may adversely affect an individual's achievements in schoolwork or in his social and emotional adjustment. Hence, in some cases they might play a sig-

nificant part in clinical work and in vocational and educational guidance. Tests are available for visual acuity, color vision, and auditory acuity.

Color Vision Tests. All such tests depend upon the principle that color-deficients confuse certain groups of hues, *inter se*, while a normal person distinguishes them. Thus, one set of charts is so devised that persons with unimpaired color vision should see certain bars, or arms, radiating from the centers of the circles. In one of the circles, for example, a person having unimpaired color vision will see two radiating arms: one green and one red. A red-blind eye will see only the green, a green-blind eye only the red, and the red-green blind eye will see neither. Another set of charts is so devised that one with normal color vision will see certain numerals, whereas the color-deficient will not. The weakness of these tests lies in their requirement of a standard illuminant for testing—a condition which is almost never in effect. For example, in the Navy, during World War II, about fifty percent of all color-deficients remained undetected after one to five medical examinations, and despite instructions to examiners to exercise great care as to illumination. However, new research is now complete on a color vision test which remains diagnostically stable despite variations in illumination.¹

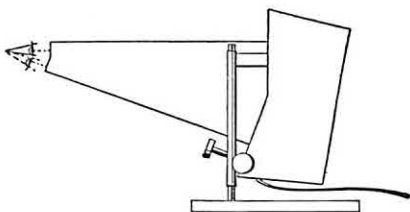
Tests of Visual Acuity. These have been taken by nearly everyone, the most familiar being the crude Snellen Chart. On this chart are printed rows of letters, varying in size, to be read by the subject. Each row and size has been standardized as recognizable at a specified distance by the "normal eye." Visual acuity is expressed as a fraction: the numerator is the distance the subject stands from the chart (usually 20 feet), and the denominator is the "distance value" of the smallest letter that can be read by the person being tested. "Distance value" of a given size is the distance at which a letter of that size can be read by the "normal eye." Thus, if the smallest letter read by a person standing at a distance of twenty feet is read by the "normal eye" at forty feet, that person's vision (in that eye) is given as 20/40. The present Snellen Chart, though still used, is a very inadequate

¹ E. Freeman, "An Illuminant-Stable Color Vision Test, I," *Journal of the Optical Society of America*, Vol. 38, 1948, pp. 532-538. This test is distributed by The Psychological Corporation.

test; for it will detect only the myopes, but not the hyperopes and presbyopes, nor those heavily handicapped by muscle imbalance.

For some years, the only available device for more thorough non-clinical testing of near and distance visual functions was the Keystone "Telebinocular," the accuracy and dependability of which have been seriously questioned. Just recently, however, a new, much more efficient device, known as the Protometer, has been developed. It is administerable by nonclinical persons for testing both near and distance visual functions (acuity, muscle balance, depth perception, and color).²

FIG. 12.1. The Protometer, designed and developed by Ellis Freeman. (Patented.) (By permission.)



The Protometer is designed for rapid and comprehensive testing of vision where numerous individuals are involved, as in schools, colleges, industries, and the armed forces.

The Protometer gives, among other data, monocular acuity, binocular acuity, and muscle balance, both for distance and for near—all under proper conditions of illumination and viewing maintained at a constant level. The Protometer discloses cases of serious impairment of vision where the trouble is not with acuity but with the failure of the two eyes to work in coordination.

The speed of operation of the Protometer is due to the fact that it is basically two Brewster stereoscopes: the optical system of the one for distance over the optical system of the one for near. As long as the subject is taking the distance test, it alone is illuminated. When the target for the near test is to be viewed, the target light for the distance test is automatically extinguished while the target light for the near test is automatically turned on. The tester operates only one control knob, which presents in sequence the test series for both distance and near and at the same time controls the illumination. The second knob is merely for adjusting the eye-height of the ocular to the subject, when this is occasionally necessary; the entire instrument is raised or lowered on a rack or pinion.

The Protometer, weighing 10 pounds, is easily portable and can be used wherever it may be plugged into current. The operator is a teacher, nurse, clerk, or other nonprofessional person who is taught to operate the instrument, read the questions from the record card, and enter responses. The record card itself informs the operator of the results, which may indicate

² Distributed by Freeman Technical Associates, 1206 Benj. Franklin Dr., Sarasota, Fla.

either satisfactory vision or deficient vision requiring referral for professional examination. The Protometer test is administered within two minutes per subject.

The test material of the Protometer consists of the basic tests used in professional refraction and has the same high precision. For this reason it is extremely dependable as a preliminary diagnostic device. It is not equipped with means for complete diagnosis and prescription of correction. For these latter a full-scale professional referral is necessary.

Since examinations of men in the last war have shown that almost ten percent of all males are color-deficient in some degree, it seems desirable to test all school children very early for this function. Deficients could, for example, be diverted from trying to become artists, geologists, clothes designers, etc. It is wise, also, to use a dependable color test in personnel divisions of department stores and of some industries; in the former to avoid placing color-deficient sales persons in the wrong departments; in the latter to avoid placing a colorblind individual on, say, radio or other wiring which requires discrimination of color code. For a job in which some color deficiency can be tolerated, the demands of the job in regard to color discrimination should be determined, as well as the candidate's degree of deficiency.

Good vision as such is desirable without question. Hence, in schools and industry there should be screening by means of a reliable device to find those who need correction or whose color deficiencies need to be given consideration in education.

Auditory Acuity. This sensory function is commonly measured (1) by means of whispered words testing the hearing of consonants and vowels, such as the Andrews Whispered Speech Test, or (2) with an audiometer. The former consists of one hundred numbers which are whispered at a specified distance. A person's score is the percentage of numbers correctly heard, divided by the "normal" percentage. The audiometer provides stimuli which are superior to whispered speech; for they are of measured intensity and unaffected by acoustics of the room in which the test is being given. The device consists of a disc and phonograph with magnetic reproducer that presents numbers of two or three digits each. The subjects hear the numbers through headphones and write them down, each ear being tested separately. Successive numbers are reproduced at decreased intensity, at small uniform steps, until a minimum intensity, scarcely audible to the "normal ear," is reached. The cycle from maximum to minimum intensity is repeated four times for each ear. Of the eight series of numbers, four

are spoken in a male voice and four in a female's. There are tests, also, of sentences, words, and pure tones.

It is obvious that these tests of vision and hearing do not measure a person's aptitude for specific types of learning and activity. For certain kinds of learning and activity, however, a given degree of visual or auditory acuity is essential. In that sense, then, these devices may constitute a part of a battery of tests which, taken together, are used to measure a particular aptitude.

MOTOR AND MANUAL TESTS

Tests of Strength of Grip. One of the oldest instruments for the measurement of individual differences in the psychological laboratory is the hand dynamometer for measuring *strength of grip*. The instrument consists of an inner and an outer handle, a dial, and a pointer. The subject grips these handles so that the second phalanges of the fingers press against the inner handle, while the outer handle presses against the heel of the hand. The subject then squeezes as hard as possible. Strength of grip is measured in kilograms. After many experiments, it appears that in psychological work this instrument is useful principally as one device for determining degree of handedness and rate of fatigue. Since these two traits are involved in certain activities and occupations, they are relevant in some aspects of aptitude testing.

Tests of Reaction Time. Reaction time is the time interval between the onset of a stimulus and the beginning of the person's overt intentional response. The particular stimulus and response to it are prearranged in an experimental situation. For example, the subject may be instructed to tap a telegraph key immediately upon perceiving a red light, the elapsed time between stimulus and response being electrically recorded in terms of thousandths of a second. It is possible to devise a variety of tests, their particular character depending upon which sensory and motor functions are to be measured. This type of test obviously is intended to measure speed of response in situations demanding immediate reaction, as in certain machine operations and in driving an automobile.

Tests of Manual Dexterity. In order to achieve competence in activities requiring manual dexterity, speed of gross movements of hand and arm, manual rhythm and coordination, and finger control and

coordination are necessary in varying degrees. For each of these purposes several tests have been devised which vary in detail but are fundamentally alike. Gross movements of hand and arm may be measured in terms of speed with which the subject picks up and places cylindrical blocks in holes in a board. Finger dexterity and coordination, necessary in rapid and accurate manipulation of objects, may be tested by measuring the rate at which an individual, with fingers or tweezers, is able to pick up small cylindrical metal pins or wooden pegs, of different shapes, and place them in the holes of a tray. (See Fig. 12.2.) Hand precision is measured by the accuracy with which



FIG. 12.2. The plier dexterity test shown here is useful in evaluating skill in the use of small tools and, in general, in evaluating aptitudes involving finger dexterity. The tray contains metal pins which must be placed in the small holes in a prescribed order. The score is based upon time required to complete the task. Sometimes the time required to remove the pins is also included in the score. (Acme Photo.)

a metal stylus can be placed into holes of small diameter cut in metal and electrically connected. Contacts of the stylus with rims of holes are electrically recorded and constitute the measure of inaccuracy. Occasionally, also, a paper-and-pencil test includes tasks designed to measure hand precision, such as speed and accuracy of tracing a path, speed of tapping, and placing a prescribed number of dots within a small circle.³

Other tests of manual dexterity follow the same general form, but some are more complex. For example, the *Small Parts Dexterity Test*⁴ consists of a metal plate having two sets of 36 holes each (six rows and six columns). One set of holes is threaded, while the other is smooth. The testee uses forceps to place a pin in each smooth hole, then a collar over the pin. In the threaded holes, the testee places a small screw, then tightens it down with a screwdriver. The stated

³ For example, the MacQuarrie Test for Mechanical Ability. Published by California Test Bureau, 1943.

⁴ By J. E. and D. M. Crawford, The Psychological Corporation, 1949.

purpose of this test is to measure a combination of perception and dexterity, in terms of rate of performance.

*The Stromberg Dexterity Test*⁵ also is a device the purpose of which is more complex than the simple measurement of manual dexterity. It consists of a tri-colored formboard (6 rows and 9 columns), into which flat, cylindrical disks, variously colored, are to be placed by the subject in a prescribed order. It appears that this test involves not only manual dexterity, but also gross color perception and a rather elementary level of nonverbal classification.

Since manual dexterity scores on tests of this type are affected, in varying degrees, by the subject's lateral dominance (the preferred use and superior performance of one side of the body or the other), it is often desirable to use tests of hand and eye dominance. This procedure is particularly indicated if we are concerned primarily with analyzing and understanding the person being tested, rather than with making selections from among candidates for a particular job.⁶

Coordination and rhythm of hand movements have been tested by means of a card-sorting test in which the subject uses one hand at a time or both hands together in dropping playing cards through slots. A more recent device is the two-hand coordination test in which the individual attempts to move both handles of a mechanism simultaneously in such a way as to keep an upper disk over the lower one, which moves in an unpredictable manner.⁷ Another two-handle device is employed in testing a subject's ability to follow an irregular path without touching the sides.⁸

During World War II, numerous psychomotor tests were used by army and navy psychologists to assist in the selection of men for specific types of training, especially in the air forces.⁹ These tests involved more difficult operations than those described above, often requiring rapid and complex sensory-motor coordination, such as the

⁵ By E. L. Stromberg. The Psychological Corporation, 1951.

⁶ See, for example, A. J. Harris, Tests of Lateral Dominance; W. R. Miles, The A-B-C Vision Test. The Psychological Corporation.

⁷ A. W. Melton, "The Selection of Pilots by Means of Psychomotor Tests," *Journal of Aviation Psychology*, Vol. 15, 1944, pp. 116-123.

⁸ For detailed summary see G. K. Bennett and R. M. Cruikshank, *A Summary of Manual and Mechanical Ability Tests*, New York: The Psychological Corporation, 1942.

⁹ Staffs, Psychological Research Unit No. 2, and Department of Psychology, School of Aviation Medicine, "Research Program on Psychomotor Tests in the Army Air Forces," *Psychological Bulletin*, Vol. 41, 1944, pp. 307-321.

following: the use of both hands simultaneously in manipulating two lathe-type handles to follow a target which moves in an irregular path; obtaining patterns of lights by manipulating stick and rudder in a simulated airplane cockpit; reacting to four different relative positions of a red light and a green light by pushing one of four switches arranged in a square pattern before the subject; moving a wheel, re-

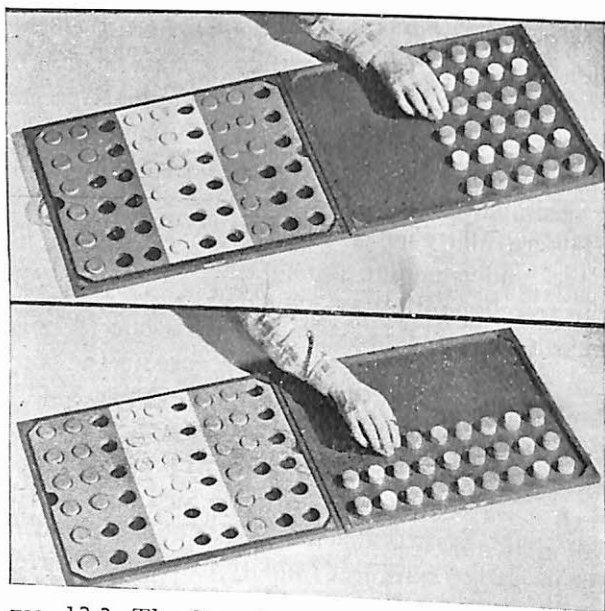


FIG. 12.3. *The Stromberg Dexterity Test.* The Psychological Corporation.

sembling an airplane control, in and out of its shaft in order to hold a horizontal bar in the center of a circular aperture; causing a beam of light to follow a given course when the horizontal movement is controlled by one lever and the vertical by another.

The tests of sensory capacity and those of motor and manual dexterity developed prior to and during World War II have been moderately useful in connection with selecting persons for specific types of training or for particular jobs. The functions and activities measured by these sensory and motor tests are, it appears, practically unrelated to the mental functions measured by tests of general ability; for the many correlational studies made between sensory and motor tests, on the one hand, and tests of intelligence, on the other, have

yielded coefficients which are very low, some being so low as to be negligible. In fact, on occasion a low negative coefficient has been found. It has been concluded, therefore, that these two types of psychological instruments measure functions which are largely independent of each other.

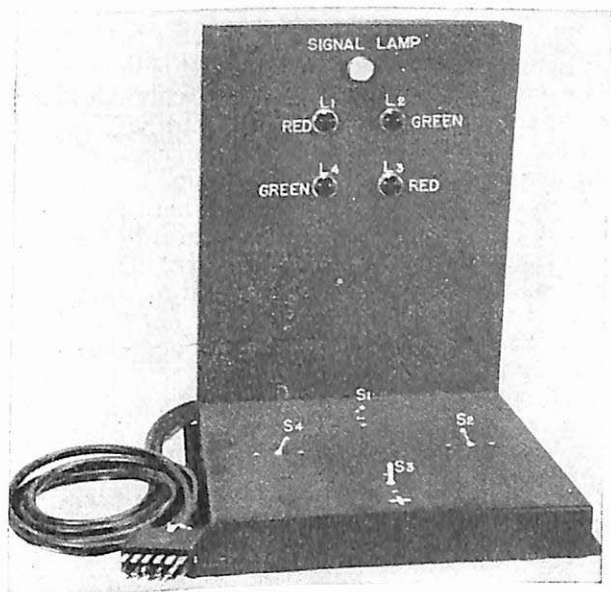


FIG. 12.4. The Discrimination Reaction Time Test. This test was designed to measure how quickly individuals make differential manual responses to visual stimulus patterns differing from one another with respect to the spatial arrangement of their component parts. The test requires that the candidate react by pushing one of four toggle switches in response to the lighting of a red and a green signal lamp. The position of the red lamp with respect to the green determines which of the four switches should be pushed.

A front view of a single test unit, with designations of lights and switches, is shown in the figure. The four stimulus lamps, two red and two green (L1, L2, L3, L4), are arranged in the form of a square on the vertical panel facing the candidate. The stimulus to which the candidate must react by operating one of the four toggle switches is the simultaneous lighting of one of the red lights and one of the green lights.

If he operates the correct switch, the white signal lamp (which lights on every trial) is extinguished immediately, signaling the candidate that he has made the correct response. The colored lights do not go out until they have been on for 3 seconds, regardless of how quickly the correct switch has been pushed.

The four spring-return toggle switches (S1, S2, S3, S4) are so set that the candidate must push each one in a different direction. The four directions of movement correspond to the four signal patterns formed by the lighting of the red and green lamps. Thus, if L1 and L4 are lighted, the red is "up" with respect to green, and the upper switch, S1, must be pushed up. If L3 and L4 are lighted, the red is to the right of the green, so the switch on the right, S2, must be pushed to the right. The time taken to operate the correct switch on each of a series of test trials is accumulated on an electric stop-clock and constitutes the candidate's score.

(From Apparatus Tests, Report No. 4, Army Air Forces Aviation Psychology Program, edited by A. W. Melton. U. S. Government Printing Office, 1947.)

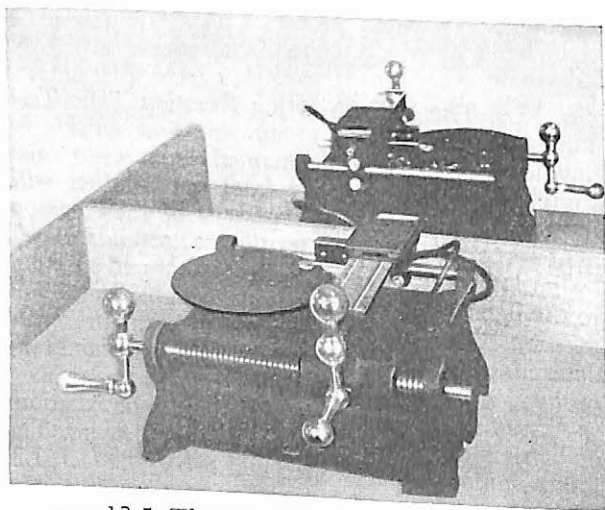


FIG. 12.5. The Two-Hand Coordination Test.

The Two-Hand Coordination Test was designed to measure a candidate's ability to coordinate the movement of both hands. He is required to control

the movements of a target-follower in response to a visually perceived target moving at varying rates along an irregular pathway.

A single test unit, as seen from the candidate's position, is shown in the figure. Two handles which he manipulates are seen in the foreground and at the left. Rotation of the upper handle causes a contact point, which is mounted on the leaf of a micro-switch, to move toward the candidate with counterclockwise rotation and away from the candidate with clockwise rotation. Rotation of the lower handle in a counterclockwise and clockwise direction causes the contact point to move to the left and right, respectively. Rotation of both handles simultaneously causes the contact point to move in any desired direction in the plane of movement of the target. A candidate's task is to manipulate the controls in such a way as to keep the targetfollower on top of a round brass button (the target) as it moves along an irregular clockwise path. When the contact point is on the target button, the microswitch is closed and current flows to an electric clock located on a remote control desk. The time which is accumulated on the clock during a series of eight 1-minute trials indicates the efficiency of the candidate's performance.

(From Apparatus Tests, Report No. 4, Army Air Forces Aviation Psychology Program, edited by A. W. Melton. U. S. Government Printing Office, 1947.)

TESTS OF MECHANICAL APTITUDE

The capacity designated by the term mechanical aptitude is not a single, unitary function. It is a combination of sensory and motor capacities such as those already briefly described, plus perception of spatial relations, the capacity to acquire information about mechanical matters, and the capacity to comprehend mechanical relationships. Thus tests of mechanical aptitude are designed to measure capacity and performance on a higher level of organization than are those of sensory-motor capacity and dexterity.

*The Assembly Test of General Mechanical Ability*¹⁰ devised by J. L. Stenquist (1923), the first of its kind and now of little more than historical interest, was intended to measure a person's ability

¹⁰ C. H. Stoelting Co., Chicago.

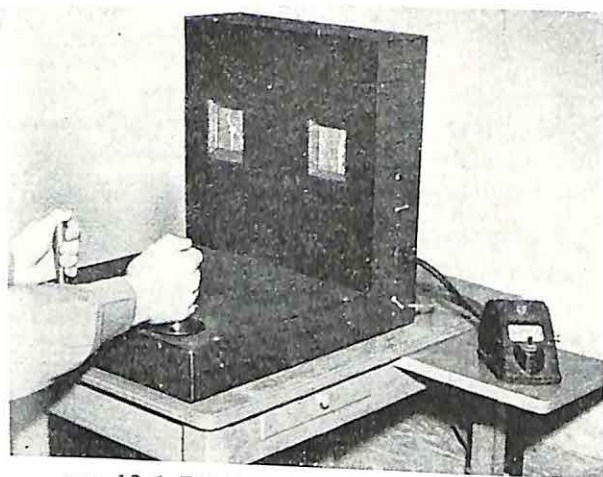


FIG. 12.6. Bi-Manual Planned Pursuit Test.

The Bi-Manual Planned Pursuit Test was designed and developed to measure ability to coordinate the activities of both hands by a systematic shifting of attention. The test consists of an irregular polished brass pathway which moves beneath two pointers. The pointers are separated by a distance of 8 inches. The pointers are adjustable by the candidate by means of two vertical handles, the candidate being required to keep the pointers (one with each hand) in contact with the moving pathway. In view of the fact that a limited amount of the pathway is visible prior to reaching the contact pointers, it was believed that a certain amount of planning could occur and would in part determine the score on the test. The test consists of six 1.5-minute trials. Rest periods of unspecified duration, probably about 30 seconds, occurred between trials. The score is the length of time during which both pointers are on the pathways.

(From Apparatus Tests, Report No. 4, Army Air Forces Aviation Psychology Program, edited by A. W. Melton. U. S. Government Printing Office, 1947.)

to put together the parts of mechanical devices, among them a bicycle bell, a double-action hinge, a door lock, and a mousetrap. This test, consisting of three series, is constructed for use with individuals covering the age range from children in the lower grades through adulthood.

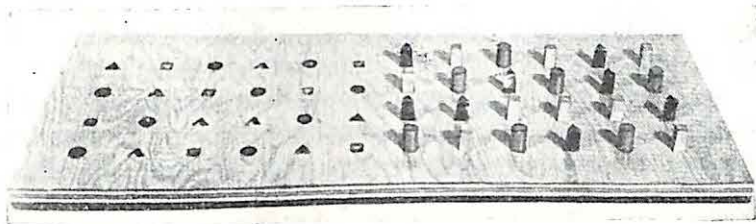


FIG. 12.7. Triform Pegboard Test.

Each test apparatus consists of two pegboards, 11.25 inches wide and 28 inches long, each containing 48 holes in 12 columns of 4 holes each, with a depth of $1\frac{1}{16}$ inch. Figure 12.7 is a photograph of one of these peg boards. There are 16 holes of each of three shapes: round, square, and triangular. Corresponding to these holes are 16 pegs of each shape. The 16 round pegs are painted red and have a diameter of $\frac{7}{16}$ inch; the 16 square pegs are painted yellow and have sides $\frac{7}{16}$ inch in length; the 16 triangular pegs are painted blue and have $\frac{1}{2}$ inch sides. All pegs have the same length, which is 1 and $1\frac{1}{16}$ inches. In one of the peg boards the various-shaped holes are scattered randomly throughout the board, while on the other board they are grouped in three sections; the round holes on the left third of the board, which is painted red; the square holes in the center section, which is painted yellow; and the triangular holes in the right section, which is painted blue.

The boards are placed in front of candidate. The board in which the holes are in irregular order contains the pegs and is placed above the board in which the holes are grouped by shape. The task of the candidate is to transfer the pegs in a standard order from upper board to lower board.

The test is designed for administration in a 15-minute test period, one minute of which is required for preliminary instructions. There are six 40-second trials, with 60-second intervals between trials for replacing the pegs in their original positions and for rest. The first three trials are performed with the right hand, and the last three trials with the left hand. There is no practice trial. Scores are recorded for each trial. The score is the number of pegs placed in the bottom board during the 40-second period. No credit is given for a peg in the candidate's hand at the signal "Stop." Credit is given for all pegs placed even though in improper order.

(From Apparatus Tests, Report No. 4, Army Air Forces Aviation Psychology Program, edited by A. W. Melton. U. S. Government Printing Office, 1947.)

The Stenquist tests have been revised and extended at the University of Minnesota (1930) and are known as the *Minnesota Me-*

chanical Assembly Test.¹¹ In principle, these are essentially the same as Stenquist's tests, some of the same mechanical devices having been retained, with new ones added. Performance on these tests—scored in terms of rate and accuracy of work—has been found useful in predicting success of junior high-school boys in shop courses. Also, facility in these assembly tests has been found by some investigators to be one significant indication of a person's aptitude for a number of occupations such as machinist and auto mechanic.

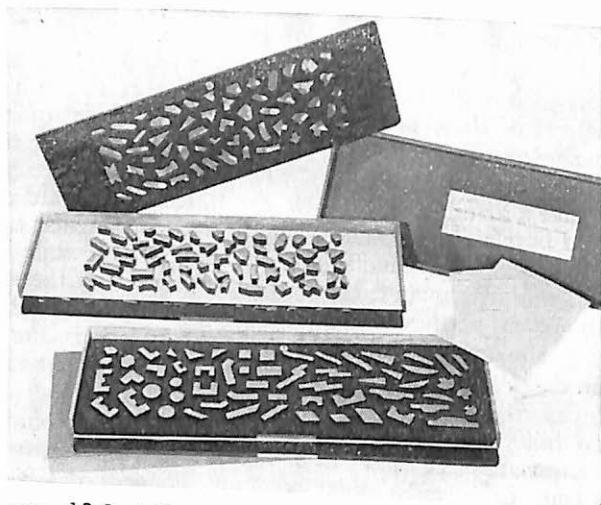


FIG. 12.8. Minnesota Spatial Relations Tests. The upper and lower parts represent the formboards which are filled with the pieces represented in the middle part. Educational Test Bureau. (By permission.)

The Minnesota Spatial Relations Test (1930) consists of a series of four boards, each of which has 58 cutouts of various shapes, many of them unusual.¹² The subject's task is to replace these in their correct holes in the board. Evidence indicates that persons engaged in mechanical occupations tend, as a group, to earn higher scores than do persons in nonmechanical occupations. This fact, it appears, is a principal justification for use of the test as a measure of mechanical aptitude. Some critics of the test have concluded that it is adequate

¹¹ D. G. Paterson, et al., *Minnesota Mechanical Ability Tests*, Minneapolis: University of Minnesota Press, 1930.

¹² Marietta Apparatus Co., Marietta, Ohio.

as a measure of speed and accuracy in responding to details of spatial relations and that it yields a measure of an individual's capacity to work with a variety of details in handling objects and concrete materials. On the other hand, it is not adequate for measuring resourcefulness in solving problems of a mechanical nature, nor for measuring capacity to manipulate small objects with precision.

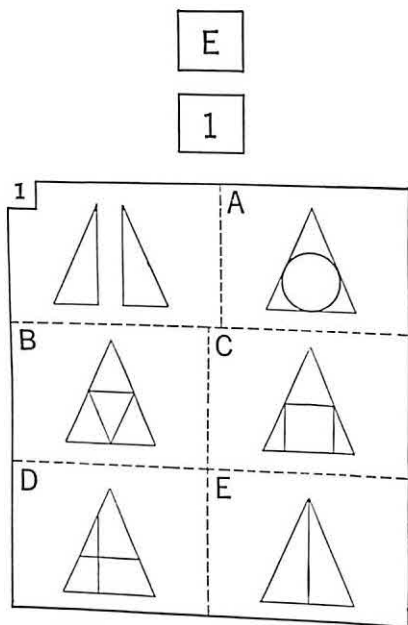
The Revised Minnesota Paper Formboard (1948) is, as its name indicates, a test which reproduces in printed form the same type of problems as those presented by actual formboards.¹³ In each problem, the subject is shown two or more parts of a geometric figure; when correctly assembled, the parts will make the complete figure. It is the subject's task to identify the correctly assembled figure from among five choices. This test is designed, it appears, to measure one's capacity to visualize and imaginably manipulate geometric forms. The data reported for it indicate that it has value for the prediction and measurement of mechanical ability and for differentiating between various grades of proficiency therein. Reported research has shown this paper formboard test to have a moderately good correlation with quality of mechanical performance and a moderate to low correlation with success in mechanical drawing and descriptive geometry. As a group, students in engineering and mechanical vocations obtain higher scores than do other groups of students. Available evidence, however, demonstrates that this test does not have high enough predictive value to be used exclusive of other criteria and information.

A number of "pencil and paper" tests are designed to evaluate mechanical aptitude by testing for specific mechanical information, selected vocabulary, and ability to perceive and deal with practical mechanical problems. One of the earliest of these is the *Stenquist Mechanical Aptitude Test* (1921).¹⁴ One part consists of problems presented by means of pictures. In each instance, the subject is required to determine which of five pictures belongs with each of five others. This part is, essentially, a test of the subject's knowledge about mechanical tools, objects, and devices, although there is some room for the perception of relations and for reasoning. The second part of the Stenquist test consists of some material similar to that in the first section (that is, matching missing pieces with the correct mechanical

¹³ Psychological Corporation, New York.

¹⁴ World Book Co., Yonkers, N. Y.

objects) plus questions applied to cuts of machines and machine parts. One underlying assumption of a test like the Stenquist must be that knowledge about mechanical tools, objects, and devices reflects mechanical interests and that mechanical interests are indicative, in some degree, of mechanical aptitude.



"First look at Problem 1. There are two parts in the upper left-hand corner. Now look at the five figures labelled A, B, C, D, E. You are to decide which figure shows how these parts can fit together. Let us first look at Figure A. You will notice that Figure A does not look like the parts in the upper left-hand corner would look when fitted together. Neither do Figures B, C, or D. Figure E does look like the parts in the upper left-hand corner would look when fitted together, so E is PRINTED in the square above **1** at the top of the page."

FIG. 12.9. Specimen Item from Revised Minnesota Paper Form Board Test. Psychological Corporation. (By permission.)

The *Tests of Mechanical Comprehension*¹⁵ present mechanical problems in pictorial form. In each instance, accompanying the picture is a statement of the problem depicted, with two or three answers from which to choose the correct one. These tests, on three levels of difficulty, are designed to measure one's understanding of the operations of physical and mechanical principles in relatively simple situations. One form is designed for use with high-school students, engineering-school applicants, and, in general with relatively untrained and inexperienced persons. A second and somewhat more difficult form is intended for use with engineering-school candidates, appli-

¹⁵ By G. K. Bennett et al., New York: The Psychological Corporation, 1940-1951.

cants for technical courses or for employment in mechanical jobs. The third form was devised for use with high-school girls and women. Since the types of items included are intended to be appropriate to the level and experience of each group of examinees, many of the items used for the women's test come within their range of household activities, involving objects and devices used in a home rather than in a shop. Other items, non-household in character, are also utilized.

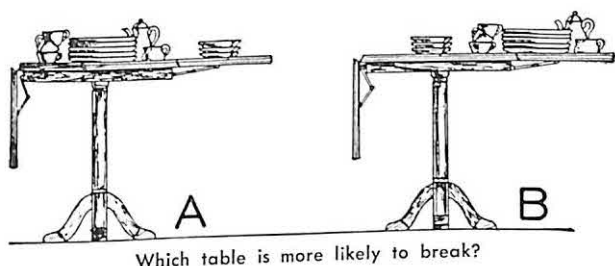


FIG. 12.10. Specimen Item from Test of Mechanical Comprehension by G. K. Bennett and D. F. Fry. Psychological Corporation. (By permission.)

Unlike some other tests of mechanical comprehension, this one does not require specific knowledge, such as matching the parts of a tool or some other mechanical object; nor does it require verbal knowledge regarding tools, processes, or materials. The items in the present test depict objects that are almost universal in American life, such as airplanes, ladders, stairs, wheels, gears, pulleys, see-saws, and others. The hypothesis is that answers to the problems presented do not depend upon specific information or training but can, rather, be arrived at by analysis of the nonverbal materials presented. The extent to which this hypothesis is satisfied varies somewhat among the 60 items of each test. For example, there is no doubt that familiarity with elementary physical principles or actual experience will be an aid in answering questions involving pulleys and leverage. Yet, individuals without such advantages, whose analytical ability is adequate, are often able to arrive at correct answers.

The Prognostic Test of Mechanical Abilities (1947) is intended for guidance purposes with pupils in grades seven through twelve and for screening purposes in industry. It is so devised as to provide a profile of abilities that are common to many mechanical occupations.

It also provides a composite score. The content of the test was determined by analyses of courses of study and by job analyses in mechanical and related occupations.

The subtests are the following: arithmetic computation, from simple addition through fairly simple fractions (no arithmetical-problem solving); reading simple drawings and "blueprints"; identification and use of common tools; spatial relationships (the paper form board); checking measurements.¹⁶

Evaluation of Mechanical Aptitude Tests. It is evident from the descriptions of the foregoing tests that mechanical aptitude is best regarded as a complex of several functions in the measurement of which some test authors emphasize only one or two aspects while others have made their devices more comprehensive.

On the whole these tests are statistically reliable. Their validity, however, may be questioned; for in evaluating them, psychologists repeatedly comment on the inadequacy of validating criteria. If we regard marks in high-school shop courses, scores of occupational and educational groups (mechanic versus nonmechanic), and low correlations with tests of general intelligence as criteria, then we can say that some of the available tests in this field have a fair degree of validity for purposes of educational guidance. On the whole, by comparison with tests of intelligence—such as those described in earlier chapters—available tests of mechanical aptitude are inferior in respect to definition of functions to be measured, level of standardization, and predictive value in actual performance.

The Wrightstone and O'Toole test—one of the more recent—may be cited for illustrative purposes. Its reliability is high, being about .90. Intercorrelations of subtest scores range from .30 to .70, with a median of .55. Correlations of subtest scores with total scores vary from .52 to .77, with a median of .70. Evidence of validity is offered in two ways: in part in terms of "face validity" and in part in terms of correlations with instructors' ratings in a training course for aviation mechanics. In the first instance, the test's authors state that their device measures the skills prescribed as necessary in eleven mechanical occupations. In the second instance, the coefficients of contingency go from .60 to .78, with a median of .67. These coefficients are among

¹⁶ By J. W. Wrightstone and C. E. O'Toole. Published by California Test Bureau, Los Angeles.

the highest in this area of investigation, and considerably higher than most others.

Numerous studies have been published in which various tests of mechanical aptitude have been intercorrelated. The reported coefficients are almost uniformly low (below .50) or very low. A few of the reported coefficients are moderate; that is, somewhat above .50. The reasons for these relatively low coefficients—unlike those found between the sounder tests of intelligence—are to be sought in the following factors. (1) Some of the tests are much more comprehensive in scope than others that are relatively restricted and homogeneous in content; hence, the former measure a greater number of functions, some of which may have little communality with the latter. (2) Not all of these tests are calibrated for the same levels of difficulty; hence they do not have equal differentiating value at a given level. (3) Some of the tests are much more dependent upon experience and specialized information than are others. (4) Performance levels on several tests may reflect different degrees of interest and motivation in special areas.

In connection with (1), above, study of the content of tests of mechanical aptitude shows that they sample, more or less, the following functions: visual-motor integration, spatial visualization, perceptual speed, manual dexterity, and visual insights (analysis). In addition to these, some of the tests measure specialized information, knowledge of techniques, arithmetical problem-solving ability, and technical vocabulary. Some of the functions are measured by means of apparatus tests (Figures 12.4, 12.5, 12.6), others by means of performance-type materials (formboards, etc.), and still others by means of pencil and paper tests. It is not surprising, therefore, that intercorrelations between these tests are low, even though they fall within the same category.

On the whole, tests of mechanical aptitude show very moderate or only low correlations with actual job performance. This fact does not necessarily signify that the tests themselves are defective, for many non-mechanical factors enter into the job ratings received and into actual performance on the job. These factors include subjective judgments of the raters, the worker's health, motivation, and personality traits which may facilitate or impede performance.

Although the foregoing factors might lower the correlation coefficients, it is highly improbable that they are solely accountable for

the relatively low relationships found. It is essential, therefore, that a given test of mechanical aptitude be studied for each type of occupation where it is to be used, in order to establish its validity not only in terms of correlation coefficients but—and more significant—in terms of critical cut-off scores and expectancy tables for each of several levels of test performance. (See the discussion of these techniques in Chapter 1.)

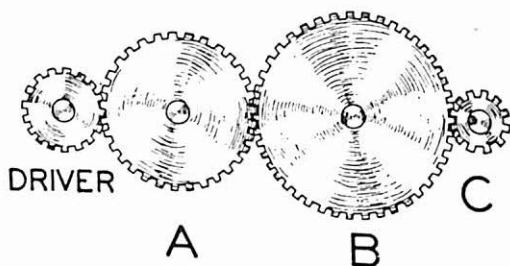


FIG. 12.11. "Which gear will make the most turns in a minute?" From the Bennett Test of Mechanical Comprehension. The Psychological Corporation. (By permission.)

Some of the tests of mechanical aptitude merit consideration when vocational or educational guidance is the problem at hand; for they are valuable as supplements to other types of information. For example, the Bennett and Fry Test of Mechanical Comprehension appears to be quite useful for selection in situations where understanding of machines is necessary. In any case, in guidance work it would be desirable to administer more than one test of mechanical aptitude, since intercorrelations of the several instruments are not nearly high enough to warrant their use interchangeably. The particular combination of tests used in any given situation should depend upon the nature of the problem presented by the individual concerned and upon the kinds of jobs under consideration.

TESTS OF CLERICAL APTITUDE

Clerical aptitude, like mechanical, is not a unitary function. The tests consist of several kinds of items, some of which correlate quite highly with scores of general intelligence tests, but differ from the latter in that they contain only selected materials which are of significance for clerical occupations.

The Psychological Corporation General Clerical Test (1950) is intended to examine routine clerical aptitude, proficiency in arithmetic, and verbal facility. The first is measured in terms of perception of similarities and differences between paired names and numbers, and in terms of skill in using a filing scheme. The second includes the usual arithmetic processes and problems. The third is tested through spelling, paragraph meaning, comprehending directions, word meaning, and language usage.

1. 9 3 7 - 9 3 7 ()
2. o b h - o b h ()
3. Curtis & Co. - Curtis & Co. ()
4.   -   ()
57. 6 8 1 9 0 0 2 3 4 1 - 6 8 3 9 0 0 2 3 4 1 ()
58. v r x o a e d i q f - v r x o a e d i q f ()
59. H. W. Hieronymous - H. W. Hiernymous ()
60.      -      ()

FIG. 12.12. Rate and Accuracy in Perceiving Similarities and Differences. From Detroit Clerical Aptitudes Examination. Public School Publishing Company. (By permission.)

The Detroit Clerical Aptitudes Examination (1944)¹⁷ attempts to measure more comprehensively than the foregoing, although the superiority of the Detroit test has not been demonstrated. The authors state that their test is intended to select pupils with capacities for commercial courses in high school. The following parts are included: rate and quality of handwriting, rate and accuracy in checking (perceiving similarities and differences between paired series of digits, letters, names, and small, simple geometric figures), simple arithmetic processes, motor speed and accuracy (placing X's in a page full of small circles), knowledge of simple commercial terms, visual imagery (disarranged parts of a picture), rate and accuracy of classification

¹⁷ By H. J. Baker and P. H. Voelker, Bloomington, Ill.: Public School Publishing Co.

(a letter-number substitution test, a variant of the digit-symbol test), and alphabetical filing.

The Minnesota Clerical Test (1946) is one of the most widely used in this area of aptitude testing. It consists of two subtests: Numbers and Names. The former consists of paired numbers, only some of which are identical and are to be marked by the testee, for example: 9632-9632; 5179-5719. The Names test is designed to measure the same functions (perception of detail and perceptual speed), using identical and non-identical paired names. The differences within the latter pairs are minor: for example, Braun and Co.-Brown and Co. These two subtests have much in common regarding psychological functions ($r = .65$); but the Names test is more inclusive and complex, for it correlates considerably higher with intelligence scales than does the Numbers test.

Evaluation of Clerical Tests. The three tests, briefly described above, will suffice as examples of this type; for they are quite representative. On the whole, results obtained with clerical aptitude tests, if critically interpreted, may contribute to a better understanding of a pupil's capacities and to his guidance in the selection of a high-school course, although the tests correlate only moderately with marks in commercial courses (from about .30 to about .50). But moderate correlations between test results and school marks are the rule rather than the exception, for the size of the coefficients is affected by several factors besides inadequacies of the tests themselves: namely, pupils' lack of interest or incentive in their courses; interference of non-intellectual factors, such as poor health, economic pressure, emotional forces, and extracurricular activities; and the variability of teachers' marks.

While their *reliability* coefficients are generally within the satisfactory range, tests of clerical aptitude do not provide sufficient evidence of their *general* value for the prediction of competence and quality of performance on the job itself. *Validity* correlations generally fall between about .20 and .45. However, as so often happens even when validity correlations are low, the tests are useful in identifying those persons at the higher levels and those at the lower. Here again, cut-off scores and expectancy tables can be more revealing than correlation coefficients.

Authors of these tests provide, among other data, norms for a variety of groups—e.g., high-school seniors in commercial courses,

office workers, non-office workers, different classes of clerical workers, employed and unemployed office workers—which indicate the trend in scores that would be expected for each of several groups. But the range of scores within groups and the extent of overlapping of scores between groups are so considerable, and correlations are low enough, so that in any individual case detailed analysis of test results must be evaluated, not in isolation, but together with other information concerning the individual under study.

DIFFERENTIAL APTITUDE TESTS¹⁸

Rationale. This battery of tests is a most ambitious and comprehensive instrument. Its statistical data on standardization and analysis—reliability, validity, intercorrelations of parts, norms, population samples—are exceptionally thorough.

Several guiding principles were applied in the development of the parts of this battery:

All eight parts of the battery were standardized on the same population. Thus the norms and percentile values for each test have the same relative significance as those for all the other tests in the battery; for the ranges of age, aptitude, school grade, and non-intellective personality factors were constant in the standardization process. Psychological profiles, therefore, are more meaningful for interpretation of differences *within* an individual. The published norms are based upon a population sample of 47,000 boys and girls in grades 8 through 12, from communities throughout the country. Separate batteries and norms are available for boys and girls.

Each test in the battery should be relatively independent and measure a relatively restricted range of aptitude.

Each test in the battery should be intended for the same purpose as every other one: namely, educational and vocational guidance (rather than, for example, selection for a particular job).

Each test in the battery should be useful for guidance in a number of related areas, rather than in only one or two.

Each test should measure *level* of aptitude—that is, “power” rather than speed of performance.

The battery of tests should yield a profile in percentile ranks. All eight percentile ranks for an individual will be comparable since they have been derived from the same population sample.

¹⁸ By G. K. Bennett, H. G. Seashore, A. G. Wesman. The Psychological Corporation, 1947. *Manual*, Second Edition, 1952.

Contents. The battery includes the following eight tests.

Verbal reasoning: verbal analogies of the familiar type are used to measure ability with more or less complex verbal concepts and relationships. Since this test encompasses only a very restricted area of verbal reasoning, "verbal analogies" would be a more appropriate name.

Numerical ability: numerical relationships and facility with number concepts are tested—that is, essentially computation rather than arithmetical reasoning (solving problems). Some of the items test only skill in the four fundamental processes; others require understanding of numerical concepts and relationships.

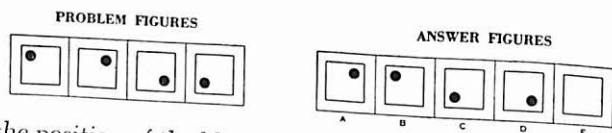
Abstract reasoning: ability to reason with nonverbal materials. The series presented in each problem requires insight into an operating principle in the changing diagrams. (See Fig. 12.13.)

Space relations: ability in spatial visualization is tested by presenting two-dimensional geometric figures (variously shaded) which are imaginably manipulated, each to form a three-dimensional figure. Using two-dimensional drawings, the purpose is to test aptitude to visualize constructed figures, variously rotated.

Mechanical reasoning: mechanical comprehension is tested with the same types of items as those already described under *Tests of Mechanical Comprehension*.

Clerical Speed and Accuracy: speed and accuracy of responses to letter and number combinations are measured. This test involves the

Each row consists of four figures called Problem Figures and five Answer Figures. The four Problem Figures make a series. You are to find out which one of the Answer Figures would be the next, or the fifth one in the series.



Study the position of the black dot. Note that it keeps moving around the square clockwise: upper left corner, upper right corner, lower right corner, lower left corner. In what position will it be seen next? It will come back to the upper left corner. Therefore, B is the answer, and you would mark your Answer Sheet like this

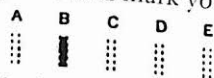


FIG. 12.13. Space problem ("Abstract Reasoning") from *Differential Aptitude Tests*. The Psychological Corporation. (By permission.)

matching of various combinations, emphasizing perception of detail and rate of response.

TEST ITEMS

V. <u>AB</u>	AC	AD	AE	AF
W. aA	aB	BA	Ba	<u>Bb</u>
X. A7	7A	B7	<u>7B</u>	AB
Y. Aa	Ba	<u>bA</u>	BA	bB
Z. 3A	3B	<u>33</u>	B3	BB

SAMPLE OF ANSWER SHEET

	AC	AE	AF	AB	AD
V				<u>.....</u>	
W	BA	Ba	<u>Bb</u>	aA	aB
X	<u>7B</u>	B7	AB	7A	A7
Y	Aa	<u>bA</u>	bB	Ba	BA
Z	BB	3B	B3	<u>3A</u>	<u>33</u>

FIG. 12.14. Number and letter test items from differential aptitude tests. In each row of the Sample Answer Sheet, the examinee marks the combination that matches the underlined combination in the corresponding left-hand row. The Psychological Corporation. (By permission.)

Language usage: Part I is simply a spelling test. Some words are correctly spelled; others are incorrectly spelled. The subject indicates for each word whether it is right or wrong. To call this test "language usage" is to stretch the meaning of that term to unreasonable lengths.

Language usage: Part II is made up of sentences in which the examinee is required to distinguish faulty from correct grammar, punctuation, and word usage.

Evaluation. As already stated, these are among the few most thoroughly standardized and reported aptitude tests available. They measure several psychological functions that have been shown to be significant in educational and vocational guidance. In this connection, the intercorrelations of the eight parts show that each has a separate and different—though not entirely unique—contribution to make; for the coefficients range (for boys) from .06 (Mechanical Speed with Clerical Speed and Accuracy) to .62 (Sentences with Verbal Reasoning). The range of coefficients for girls is approximately the same (.12–.67).

Extensive correlational studies with other standardized tests, of a variety of types, show that the several parts of the DAT do not measure entirely the same functions as do these others, although some of the parts (Verbal Reasoning and Numerical Ability) are often highly correlated with tests of general intelligence. The DAT, therefore, has a contribution to make in a comprehensive study of an individual's aptitudes.

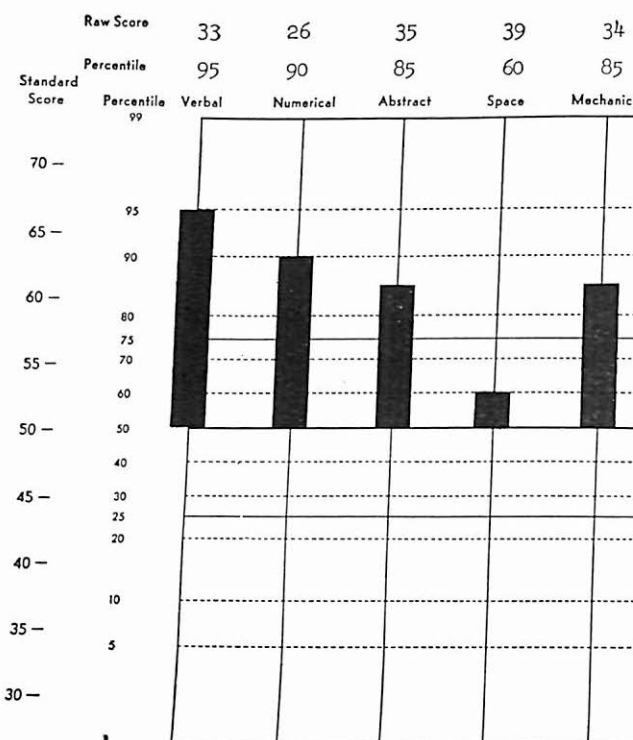


FIG. 12.15. Differential aptitude tests. A profile of scores (in parentheses.)

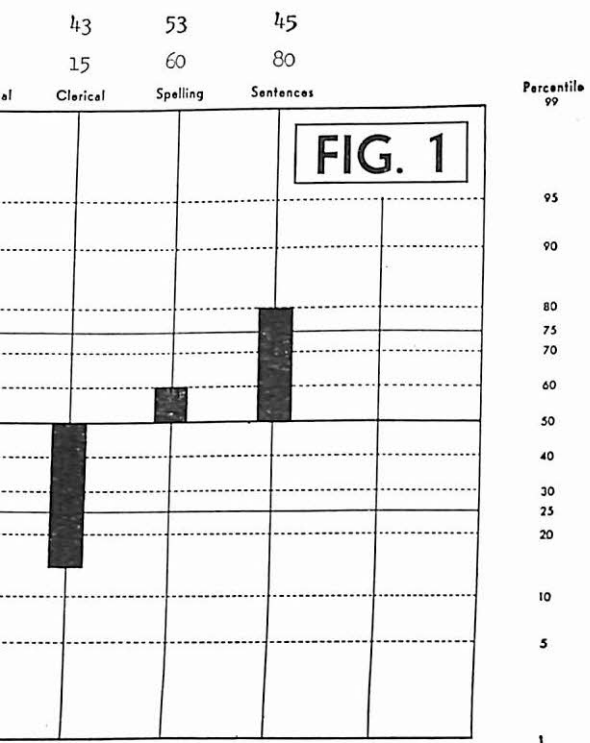
the psychological rationale presented in the Manual; and both of these will make more than the usual demands upon the non-specialist's technical sophistication.

As is the case with all tests of aptitudes, the ultimate value of the DAT will depend upon follow-up studies showing their predictive efficiency in learning and job performance. For this purpose, the authors and users of this battery of tests will have to determine the differential and predictive significance of profile "patterns" and multiple correlations (between two or more parts of the tests, on the one hand, and the criterion, on the other), thus going beyond the simple correlation coefficients and expectancy tables, important though these are.

APTITUDE CLASSIFICATION TESTS ¹⁹

This battery of tests has been devised for purposes of vocational guidance. Its recommended occupational range (thirty vocations) covers a wide area at several levels: including such different

¹⁹ By J. C. Flanagan, Science Research Associates, 1953. This test is described at this point because it is intended for use in vocational guidance, covering a range of occupations from clerical and mechanical through some professions. It may, therefore, serve as a bridge to the following chapter.



res. The Psychological Corporation. (By permis-

Aptitude Classification Tests

Extensive data are also reported showing validity coefficients between DAT scores and grades in a variety of school subjects. These indexes are of approximately the same magnitude as those found with other sound tests. The correlations on the parts of the DAT vary with the several school subjects used as validity criteria; so that, as is to be expected, some parts have greater predictive value in a given area of studies than do others.

In addition to the foregoing, norms of performance are provided for a number of different educational and occupational groups.

The tests are adequately reliable for the most part, the mean coefficients for the several grade levels varying from .85 to .93 for boys, and .71 to .92 for girls. The test of Mechanical Reasoning (reliability = .71) is the only part that is seriously below accepted degrees of reliability.

The scores of this battery are reported in profile (graph) form, as shown in Figure 12.15. The interpretation of such a profile and the significance to be attached to score differences as shown on the graph depend upon adequate comprehension of the statistical evidence and

occupations as office clerk, mechanic, humanities professor, nurse, physicist, and writer.

The battery comprises fourteen tests, each of which is briefly described below:

- inspection: ability to spot flaws
- coding: speed and accuracy in coding typical office information
- memory: recall of codes learned in coding test
- precision: speed and accuracy in making small circular finger movements
- assembly: ability to visualize the appearance of an object from a number of separate parts
- scales: speed and accuracy in reading scales, graphs, and charts
- coordination: ability to coordinate hand and arm movements
- judgment and comprehension: ability to read, understand, reason, and use good judgment
- arithmetic: the four fundamental processes
- patterns: ability to reproduce simple pattern outlines
- components: ability to identify important component parts in line drawings and blueprint sketches
- tables: reading two types of tables—one using numbers, the other using words and letters
- mechanics: understanding mechanical principles
- expression: communicating ideas in writing and talking

This device is of the differential aptitude type. Its author suggests various combinations of the fourteen tests as having prognostic value in the selection of a vocation. For example, for the prospective occupation of nurse, the three recommended tests are (1) memory, (2) scales, (3) judgment and comprehension; for the prospective humanities professor the recommended tests are (1) memory, (2) judgment and comprehension, (3) expression. So far as this battery is able to differentiate between aptitudes, then, the distinction between these two widely diverse professions appears to depend upon the tests of expression and of scales. It is hardly probable, however, that the differences in aptitudes, in this instance, are as simple and limited as these sets of tests would indicate. Nor is it probable that valid selections for either occupation alone can be made on the basis of these few and simple tests. The author of the test battery recognizes this fact, for he states that the test combinations are only tentatively proposed; and he recognizes that tests for the measurement of additional "job elements" may be necessary. Similar comparisons can be made

between many pairs of the thirty occupations for which test combinations are recommended.

So far as statistical analysis of standardization data is concerned, the results are fairly satisfactory. Intercorrelations between individual tests are generally low: seven of ninety-one coefficients are .50 or higher; the median is .29. Standard errors of measurement are moderate. Reliability coefficients for individual tests are fair (.26 to .86, Form A; .55 to .85, Form B). However, reliability coefficients for *combined* scores of all tests recommended for a particular occupation are much higher for nine occupations reported: namely, from .83 to .93. These reliability data make it clear that the "job element approach," upon which this battery is based, yields much less stable results than does a more nearly "wholistic" or "global" approach.

So far as predictive (on-the-job) validity is concerned, the value of these tests remains to be demonstrated. The standardization population consisted of high-school seniors. Only a few of these have been followed-up and rated in their occupations. The correlations between vocational test scores and occupational progress after four years (rate of salary increase), in seven occupations, ranged from .29 to .64. Correlations of test scores (obtained in senior year of high school) with college grades were from .24 to .36.

It is clear that in their present state, the Aptitude Classification Tests must be regarded as tentative efforts in the direction of vocational guidance by means of tests based on the "job element approach."

APTITUDE TESTS: FINE ARTS AND PROFESSIONS

TESTS OF MUSICAL APTITUDE

The Seashore Tests. The most widely known are the *Seashore Measures of Musical Talent* (1919-1939).¹ The six parts, recorded on phonograph records, are devised to measure (1) pitch discrimination (graded from readily discernible differences to very fine ones); (2) intensity or loudness discrimination (pairs of clicks, with gradation differences); (3) time discrimination (differing time intervals between clicks); (4) discrimination of timbre, or tone quality (pairs of tones differing in quality); (5) rhythm discrimination (pairs of rhythmic patterns of increasing complexity, to be discerned as the same or different); and (6) tonal memory (paired tonal patterns requiring perception of differences between members of the pair). There are two sets of records: one for unselected groups and the other for musicians and students of music. Total scores for the six parts do not represent an individual's test performance; instead profiles of the six auditory capacities should be used and evaluations should be based upon them.

Seashore approached the development of his measures from a point of view different from that of authors of most other types of aptitude tests. Instead of making a "job analysis," which would attempt to discriminate levels of musical aptitude on an empirical basis, Seashore made what he regarded as a theoretical analysis of musical talent

¹ RCA Manufacturing Co., Camden, New Jersey.

into its sensory components. Some of these components, he held, can be measured objectively, whereas others cannot. The six capacities mentioned above are among the measurable ones; but as Seashore himself insisted, they do not offer a complete index or profile of all components of musical aptitude. They are measures only of auditory perception required in music, analyzed into six components.

For the purpose of validation, results of the Seashore measures have been compared with teachers' ratings of musical ability, with musical achievement, and with quality of work in schools of music. The obtained correlations have not been such as to warrant the conclusion that these tests of auditory perception are sufficiently valid in predicting various levels of musical talent. Seashore himself, however, has objected to attempts at an over-all validation of his measures. While he has not done so himself, he maintains that each of the six tests should be separately validated against different kinds of specialized musical activity. For example, the test of pitch discrimination should be validated especially for players of string instruments. At any rate, this much may be said: the Seashore measures reveal those persons whose auditory perceptions are so deficient that they could not successfully participate in the formal study or performance of music.²

The Wing Tests. *The Wing Standardized Tests of Musical Intelligence* (1948),³ also presented by means of a series of recordings, is intended to meet at least two objections that frequently have been directed against the Seashore tests: namely, that the latter are "atomistic" and that they are not based upon functions being trained and considered most important by teachers of music. The Wing tests measure seven aspects of musical perceptiveness, yielding a score for

² See J. C. Saetveit, D. Lewis, and C. E. Seashore, *Revision of the Seashore Measures of Musical Talent*, Iowa City: University of Iowa, 1940; H. M. Stanton, *Measurement of Musical Talent*, University of Iowa, 1935; C. E. Seashore, *Psychology of Music*, New York: McGraw-Hill Book Co., 1938; J. L. Mursell, *The Psychology of Music*, New York: W. W. Norton, 1937; Paul R. Farnsworth, "An Historical, Critical, and Experimental Study of the Seashore-Kwalwasser Test Battery," *Genetic Psychology Monographs*, Vol. 9, 1931, pp. 291-393.

³ By H. D. Wing and C. Wing. Sheffield City Training College, Sheffield, England. Cf. O. K. Buros, *Fourth Mental Measurement Yearbook*, pp. 344 ff. The use of the term "musical intelligence" is unfortunate, since the word "intelligence" has been given another psychological connotation.

each; but test performance is represented by only a single total score, rather than by part-scores, since "musical intelligence" is held to be a unitary, though complex, aptitude.

The seven parts are: (1) chord analysis, (2) pitch change, (3) memory, (4) rhythmic accent, (5) harmony, (6) intensity, (7) phrasing. These aspects of musical capacity are held to be more closely associated with actual teaching of music than are Seashore's, since Wing's standardized tests are based upon the individually constructed and subjective types of tests employed by teachers and examiners in music. The Wing tests, therefore, are considered by many teachers of music to be more relevant to the selection and training of persons in that art.

The Drake Test. *The Drake Musical Memory Test* (1934) rests upon the premise that immediate memory of melodic materials is a basic factor in musical ability.⁴ It consists of a series of paired melodic items, the members of each pair differing in key, or time, or notes.

The Madison Test. *The Interval Discrimination Test* (1942), by Madison, is, like the Seashore tests, an example of a measure based upon a theoretical assumption.⁵ The premise in this instance is that degree of ability to make interval discriminations is a measure of musical aptitude. The intervals vary from those easily perceived to those requiring very fine auditory discrimination.

Evaluation of Music Tests. The foregoing tests of musical aptitude are reasonably reliable measures of the functions they measure, when the total scores are considered. On the Seashore, for example, reliabilities of the part-scores range from only .62 to .88 (Series A). This means that the part-scores often are not reliable enough to discriminate, except grossly, between the several sensory functions within each person. Undoubtedly, the total Seashore score is more reliable;

⁴ R. M. Drake, "Four New Tests of Musical Talent," *Journal of Applied Psychology*, Vol. 17, 1933, pp. 136-147; "The Validity and Reliability of Tests of Musical Talent," *ibid.*, pp. 447-458. Test materials are obtained from Public School Publishing Co., Bloomington, Ill.

⁵ T. Madison, *Interval Discrimination as a Measure of Musical Talent*, Archives of Psychology, No. 268, 1942. The test is obtainable from its author, University of Indiana, Bloomington, Indiana.

but since the Seashore theory posited separate factors rather than a unitary function, total scores are not reported and total score reliabilities are not given.

The Wing tests, however, are based upon the theory of a unitary function; their total score reliabilities, therefore, are reported. The coefficient of reliability, using test-retest total scores, average .95 for an interval of one year, and .86 for intervals varying from two to four years. These coefficients fall within the range of satisfactory reliabilities.

It is admittedly very difficult to validate a test of musical aptitude against a criterion of achievement, since grades in music are fragmentary or nonexistent, grading is exceedingly subjective, training opportunities are markedly unequal, and interest and motivation are highly variable from person to person. Thus it is that even after the years the Seashore tests have been in use, many psychologists believe their predictive value has not been sufficiently demonstrated. Available Seashore validity coefficients vary from low to moderate.

The Wing tests present weightier evidence of validity, including a study of above-average, average, and below-average students of musical instruments (ages 14-16; $N = 333$ boys). Forty percent of the lowest group had discontinued their musical studies, as compared with 27 percent of the average group and 2 percent of the highest group. A finer classification would have defined the failures even more precisely.

On the whole, it appears that in spite of uncertain over-all validity, these tests are helpful in identifying persons whose prospects are poor and those whose are high for profiting from instruction in music, so far as these essential psychological functions are involved.

In spite of the differences in hypotheses and scoring between the Seashore and the Wing tests, it appears from factorial analyses that the two instruments are measuring the same aptitudes to a marked degree, when the parts of each are taken in combination as Wing has done. It appears that there is a "general factor" operating in the several tests, accounting for 30 to 40 percent of the variance found among individuals. This general factor is identified as essentially a complex perception of the sensory aspects of music. Since this appears to be the case, it would be desirable to measure these basic aspects of musical aptitude by means of the more complex processes such

as are sampled by the Wing tests.⁶ This method of testing is indicated especially in view of the fact that teachers of music and others in that field maintain that the separate sensory functions (as measured in the Seashore) do not operate in music as they do in the controlled and isolated test situation.

TESTS OF APTITUDE IN THE GRAPHIC ARTS

The Knauber Test. *The Knauber Art Ability Test* (1932-35),⁷ for use in grades 7 to 16, requires the following: drawing a design from memory; drawing from memory figures within space limitations; drawing a stereotyped character such as Santa Claus; arranging within a given space a specified composition; creating and completing designs from supplied elements; spotting errors in drawn compositions, such as incorrect perspective, misplaced details, incorrectly proportioned details, incongruous or inconsistent elements; production of compositions intended to show creative imagination, ingenuity, ability to represent a concept symbolically, or to plan and execute a universal idea.

Attempted measurements of these complex capacities represent a large undertaking for a single test. Performance on some parts, it appears, is dependent upon mastery of stereotypes and traditional problems and tasks presented in art instruction. This being the case, the Knauber test would be useful primarily for evaluating school progress in art, quality of observation, and, to some extent, creative imagination. That is, the test would be in large part a measure of learning rather than a measure whose primary usefulness is with individuals who have not had the benefits of some formal instruction.

The Meier Test. *The Meier Art Judgment Test* (1940),⁸ for grades 7 to 12, consists of 100 pairs of uncolored pictures, one member of each pair being a reproduction of a masterpiece. The pictures cover a wide variety of subjects: landscape, still life, woodcuts, oriental drawing, murals, and others. The second member of the pair is altered

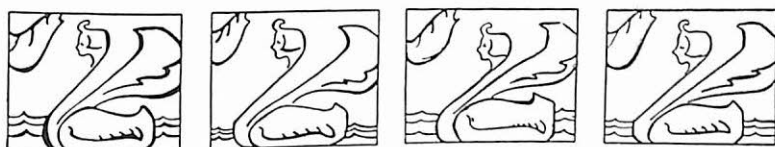
⁶ See J. McLeish, "The Validation of Seashore's Measures of Musical Talent by Factorial Methods," *British Journal of Psychology, Statistical Section*, Vol. 3, 1950, pp. 129-140. Also, H. D. Wing, "A Factorial Study of Musical Tests," *British Journal of Psychology*, Vol. 31, 1941, pp. 341-355.

⁷ Psychological Corporation, New York.

⁸ Bureau of Educational Research and Service, State University of Iowa, Iowa City.

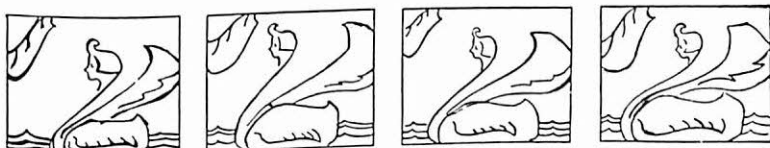
PROBLEM 3

This is a test of accuracy and observation.



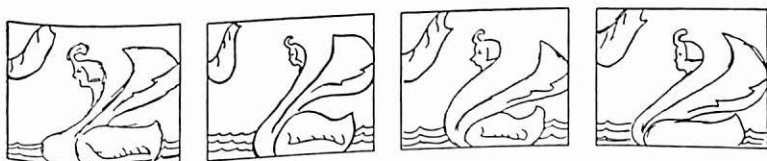
Score 10

These drawings are good in proportion and quality of line. Details are accurately observed.



Score 6

The proportions in these drawings are not as good, and the lines are weaker.



Score 3

These drawings are poor in proportion and in execution. They indicate a lack of careful observation.

FIG. 13.1. The subject is shown a design in the examination booklet which he is asked to copy. This item is a test of accuracy and quality of reproduction. From the Knauber Art Ability Test. (By permission.)

from the original in some respect and in such a manner as to make it inferior to the original. The nature of the alteration in each pair is indicated (perspective, use of curves, arrangement, etc.); the subject is, of course, not told which one is the original and which is the altered copy. It is his task to indicate his preference as between the two copies of each pair. The author of this test has selected not artistic execu-

tion but, rather, aesthetic judgment as the principal factor in art aptitude and as the only factor to be tested, the contention being that judgment is the "key-capacity," the most trustworthy and significant index to talent in art and to success in a career in art. The soundness of this assertion, however, remains to be validated. If it were correct that artistic perception and judgment constitute the "key-capacity,"

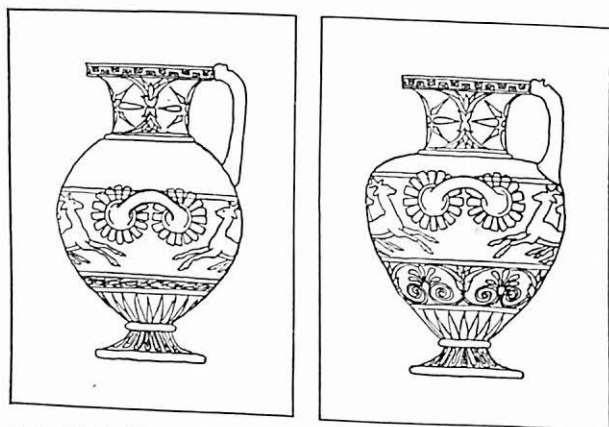


FIG. 13.2. One of these pictures represents an artistic work of established merit. The other is an adaptation of that work, and aesthetically inferior. In this pair, the subject is required to select the original and aesthetically superior work on the basis of the shapes of the bowls. From the Meier Art Judgment Test. Bureau of Educational Research and Service, State University of Iowa. (By permission.)

rather than actual execution and creativity, we should then expect art critics and writers on art to be the most talented in actual production; but such is not the case.

The Graves Test. The *Graves Design Judgment Test* (1948)⁹ is devised to measure appreciation and production of art. The author enumerates and describes eight principles that he regards as basic: namely, unity, dominance, variety, balance, continuity, symmetry, proportion, and rhythm. The purpose of the test is to evaluate the extent to which an individual is able to apprehend and respond to these principles.

⁹ Psychological Corporation, New York.

Each of the 90 items consists of either two or three plates from which the testee selects the one he regards as superior to the other one or two. The pictures in each item are two-dimensional or three-dimensional designs ("non-representational"), only one of which incorporates the principles of design upon which the test was devised. The designs are such as to eliminate or minimize subject matter and personal connotations and feelings, while emphasizing relatively pure aesthetic choice.

All eight principles are incorporated within one design in each set (item), whereas one or more principles are violated in the other designs of each set. Thus this test attempts to evaluate a unitary complex process (aesthetic judgment) rather than making separate analyses of elements that enter into aesthetic judgment. In this respect the Graves test is consistent with the Wing test in music and with the principle of a general factor in intelligence testing.

It is important to note the criteria upon the basis of which the items were selected in this complex and highly subjective fine art. Retained items in the final test satisfied all three of these criteria: (1) agreement among teachers of art as to the superior design in each set of plates; (2) greater preference for a design by students of art than by non-art students; and (3) greater preference for a design by high-scoring subjects than by low-scoring subjects. These are the criteria generally employed in this field of test construction.

Evaluation of Art Tests. The authors of tests of aptitudes in the visual arts are not in complete agreement regarding the capacities to be measured. Nor does it appear that any one of the sounder tests, such as those described above, is markedly superior to the others in predictive value. The tests are of two general types: production of specified art forms (performance), as in the Knauber, and aesthetic judgment, as in the others. Since aptitude for aesthetic judgment is not equivalent to aptitude for art production, a combination of both types should have greater selective and predictive value than either one alone for educational and vocational purposes.

It is not surprising that tests in this category differ in content, for experts in this highly subjective field of art are not in close agreement regarding the basic criteria, or principles, to be employed in the selection of superior designs or pictures, that is, regarding aesthetic judgment. It is more than ordinarily difficult, therefore, to establish uni-

versal criteria for the selection and scoring of test items and for the validation of test scores, once the items have been selected.

Closely related to the foregoing problem is the inevitable result that different teachers and critics of art, employing varying criteria in their evaluations, will rate art productions differently. The result is that there are relatively few useful data on the predictive efficiency (validity) of tests in the fine arts with regard to level and quality of performance in educational and vocational activities. Where coefficients are reported, they are only fair to low, when correlated with marks in art courses or with teachers' estimates of students' abilities. At the same time, however, available tests are useful in discovering individuals of unusual capacity and, in general, in evaluating appreciation of some of the graphic arts. These tests, also, quite consistently differentiate between art students and non-art students, the differences between their mean scores being significant.

For non-art students, the tests can be useful in evaluating capacities that are essential in aesthetic appreciation, as a basis for non-vocational education in the fine arts. This is true even though the capacities measured by the tests, though essential, are not sufficient in themselves to indicate subsequent quality of creative or even only reproductive performance.

Finally, the available tests of aesthetic judgment assume, apparently, that this capacity is generalized and transferable from evaluations of pictures and of relatively pure designs to architecture, furniture, clothing, sculpture, and industrial products. This assumption may be warranted. Actual demonstration of its validity would be of significance in school and college courses that offer instruction designed to cultivate aesthetic judgment. Definitive research has yet to be done on this psychological and educational problem.

TESTS OF APTITUDE IN MEDICINE¹⁰

General Characteristics. Several tests of medical aptitude have been constructed. The first one has appeared in 25 revised forms over a period of about twenty years, under the sponsorship of the Committee on Aptitude Tests for Medical Students of the Association of American Medical Colleges, and under direction of F. A. Moss.

¹⁰ See I. L. Kandel, *Professional Aptitude Tests in Medicine, Law, and Engineering*, New York: Teachers College, Columbia University, 1940.

Recent editions have included all or most of the following subtests: visual memory, memory for content, scientific vocabulary, understanding of printed material, scientific definitions, and logical reasoning.

These tests have been based upon an analysis of the qualifications necessary for the successful study of medicine, which are given as the following. First, it is essential to have sufficient mental alertness to learn quickly and to organize the material learned so that it can be retained and utilized in later work. A sampling of medical materials was used in the test to examine this capacity. Second, past scholastic performance may be expected to indicate future learning. Inasmuch as all premedical students have had elementary courses in chemistry, physics, biology, and English, sections of the test are devoted to questions in these subjects to determine the extent of the candidates' learning in them. Third, the capacity to make correct interpretations and deductions from given data was considered essential. Hence, the test included a passage of difficult reading, dealing with materials found in medical studies. The testee is required to make certain interpretations and deductions based upon the passage, to which he may refer at any time. Fourth, since medical students and practicing physicians are expected to draw conclusions and make diagnoses from given facts, a subtest was devised to evaluate "logical reasoning." This consists of a set of premises and conclusions drawn from them, the student's task being to determine whether or not the conclusions are warranted.

The reader has surely noted that the mental capacities to be measured by means of the foregoing types of items are by no means peculiar to the study of medicine. They are, indeed, capacities which are required in all fields of study and in all professions. The tests of this medical committee, therefore, though they are called aptitude tests, are actually tests of general ability, *utilizing in part a special content* which is included in or closely associated with the materials of study in medical schools. In other words, the *form* of mental activity being tested is the same as in any other professional field, but the *content* is in part specialized. It is in this sense that this test and others like it are aptitude tests, as that term has been defined. It is important to note this fact; for otherwise one might get the impression that a specialized aptitude, independent of general ability, is required for the study of medicine.

Contents. At present, the *Medical College Admissions Test* (1946–1951) is being further developed and administered by the Educational Testing Service¹¹ for the Association of American Medical Colleges. The purpose of this aptitude test was stated to be: to provide highly dependable measures of the advanced student's *general ability* and of his *achievement in a special field of study*. The tests are predicated upon the principle that a significant aspect of potentiality for a specialized field of study at the graduate and pre-professional level may be measured by testing the candidate's general scholastic ability and his achievement in a special field which is prerequisite to advanced study in the same or a closely related field.

In addition, a test of "understanding modern society" has been included in the battery. This part includes current history, economics, political science, and sociology, the purpose being to evaluate alertness to social issues rather than materials retained from college courses in these fields of study.

To measure *general ability*, the battery has two major divisions: verbal and quantitative. Verbal ability is measured by tests of vocabulary (word opposites), sentence completion, and word analogies. Specimens of these follow.

Opposites:

Harmony: (a) accord (b) affinity (c) oppression (d) conflict
(e) desecration

Sentence completion:

The manufacturing of small machinery is profitable for the nation lacking mineral resources, inasmuch as _____ can be exploited in place of _____.

(a) quantity-quality; (b) power-efficiency; (c) skills-quality;
(d) alloys-ores; (e) skills-materials

In addition to the foregoing, the test on current society and the test on materials from college courses in basic sciences may properly be regarded as contributing to an evaluation of the candidate's general verbal ability, although these two are not intended primarily for this purpose. The reported correlation of .77 between verbal test scores and scores on understanding modern society supports this view, as do the other intercorrelations between the nonquantitative parts (all above .65).

¹¹ Princeton, New Jersey. This test was formerly called the Professional Aptitude Test.

Ability in quantitative thinking is measured with tests of arithmetical reasoning, applied and abstract.

The test of achievement in a specialized field of study is, naturally, a science test covering a wide sampling of concepts and problems from basic college courses in biology, chemistry, and physics. These items are intended to evaluate the candidate's "grasp of fundamental principles of science."

The science test includes completions, classifications, analogies, quantitative comparisons, and paragraph comprehension. Taken as a whole, the science test is concerned with basic scientific concepts, and with applications and problems in the several fields. The problems are of increasing complexity requiring comprehension, interpretation, inference, and analysis of data—in general, the use of knowledge in dealing with multi-phase problems. Sample items and descriptions follow.

COMPLETIONS

Sample Directions: Each of the following incomplete statements is followed by five suggested completions. Select the one completion which is best in each case and indicate your selection in the appropriate space on the answer sheet.

16. A sodium atom and a sodium ion
- (A) contain the same number of electrons
 - (B) contain the same number of protons
 - (C) have the same chemical properties
 - (D) have the same physical properties
 - (E) have different atomic numbers

CLASSIFICATIONS

Sample Directions: Each of the numbered words or phrases below is associated with one, both, or neither of the headings listed as (A) and (B) above it. On the appropriate line of the answer sheet blacken the space under

- A if the numbered word or phrase is associated with (A) *only*,
- B if the numbered word or phrase is associated with (B) *only*,
- C if the numbered word or phrase is associated with *both* (A) and (B),
- D if the numbered word or phrase is associated with *neither* (A) *nor* (B).

Sample Questions:

- (A) Thyroid gland
- (B) Pituitary gland

- (C) Both
(D) Neither

17. Giantism
18. Low blood calcium
19. Cretinism
20. Short stature

ANALOGIES

Sample Directions: Each of the following questions consists of an incomplete analogy with five suggested completions. Select the one word or phrase which best completes the analogy and indicate your selection in the appropriate space on the answer sheet.

Sample Questions:

21. ohm : resistance :: watt : _____
(A) electricity (B) work (C) power (D) current
(E) potential
22. atom : molecule :: element : _____
(A) electron (B) mixture (C) isomer (D) isotope
(E) compound
23. yolk : egg :: _____ : bean seed
(A) hypocotyl (B) epicotyl (C) cotyledon (D) testa
(E) endosperm

QUANTITATIVE COMPARISONS

Sample Directions: The following paired statements describe two entities which are to be compared in a quantitative sense. On the appropriate line of the answer sheet blacken the space under

- A if A is *greater than* B,
B if B is *greater than* A,
C if the two are *equal or very nearly equal*.

Sample Questions:

24. (A) The total resistance of two given resistances in series
(B) The total resistance of the same two resistances in parallel
25. (A) The volume occupied by one gram-molecular weight of helium at standard conditions
(B) The volume occupied by one gram-molecular weight of oxygen at standard conditions
26. (A) The concentration of oxygen in the right auricle of a mammalian heart
(B) The concentration of oxygen in the left auricle of a mammalian heart

PARAGRAPH COMPREHENSION

Sample Directions: In this part of the test there are several passages, each followed by a series of statements. Read the passage and then

classify each of the statements under one of the following categories:

- (A) The statement is warranted by information given in the passage.
 - (B) The statement is true but not warranted by the passage.
 - (C) The statement is contradicted by the passage.
 - (D) The statement is contradicted by established evidence but not by the passage.
- [Scientific passages then are given, each followed by several statements to be marked as directed above.]

In a second part of paragraph comprehension, the passages are followed by five questions, from which the best answer is to be selected, based upon material in the paragraph.

Sample Passage:

The only carbohydrate which the human body can absorb and oxidize is the simple sugar glucose. Therefore, all carbohydrates which are consumed must be changed to glucose by the body before they can be used. There are specific enzymes in the mouth, the stomach, and the small intestine which break down complex carbohydrates. All the monosaccharides are changed to glucose by enzymes secreted by the intestinal glands, and the glucose is absorbed by the capillaries of the villi.

The following simple test is used to determine the presence of the monosaccharides. If Benedict's solution is added to a solution containing glucose or one of the other monosaccharides and the resulting mixture is heated, a brick-red precipitate will be formed. This test was carried out on several substances and the information in the following table was obtained. "P" indicates that the precipitate was formed and "N" indicates that no reaction was observed.

<i>Material Tested</i>	<i>Observation</i>
Crushed grapes in water	P
Cane sugar in water	N
Corn syrup	P
Molasses	N

Sample Questions on Passage:

31. From the results of the test made upon crushed grapes in water, one may say that grapes contain
- (A) glucose
 - (B) sucrose
 - (C) a monosaccharide
 - (D) no sucrose
 - (E) no glucose

32. The carbohydrate content of which one of the following foods probably undergoes the LEAST change during the digestive process in the human body?
- (A) Cane sugar
 - (B) Corn syrup
 - (C) Molasses
 - (D) Bread
 - (E) Potato

Evaluation of Medical Aptitude Tests. The purpose of these tests is to provide an improved basis for predicting quality of performance in medical *studies*, not in medical practice. In some instances, the tests have shown better results than have premedical course marks; in other instances the reverse has been true. But in all instances, Moss found that the best criterion is a combination of test results and premedical course marks. For example, he reports in one study the following correlations: premedical course marks with medical school averages, 0.67; test scores with medical school averages, 0.64; medical school averages with a combination of test scores and premedical marks, 0.81 (multiple correlation). These are very satisfactory coefficients. Other correlational studies yielded coefficients that were higher in some instances but much lower in others.

The variation in correlation coefficients between test scores and medical school grades, found in different studies, is not attributable solely, or perhaps even principally, to defects in the medical aptitude tests. The differences among coefficients must also reflect serious differences in medical school grading standards, inequalities of undergraduate preparation (which to some extent can be compensated for in medical-school courses), and personality traits which tend to produce inconsistencies between promise and performance.

Expectancy statistics have also provided useful findings regarding test scores. For example, one study reports that only one percent of the highest-decile students failed in medical school, whereas eighteen percent of the lowest-decile students failed. These findings provide an argument for admitting all top-decile students but *not* for refusing admission to those in the bottom decile group. Another study reported that the lowest decile group contributed 25 percent of the failures—that is, two and a half times its quota, in proportion to the total group of students.

Reliabilities of the *Medical College Admissions Test* fall well

within the acceptable range, namely, from .89 to .94 for the several parts. These reliability data again show that it is not as difficult to establish reliability as it is to demonstrate validity.

While these aptitude tests are not intended to predict effectiveness in medical practice which, like other professions, is dependent upon a complex of factors, the tests appear also to have some value in forecasting medical students' levels of success in internships. When a group of interns were rated on a five point scale by their hospital staffs, the results showed that the tests have some selective value, especially in identifying students who prove to be the most satisfactory interns.¹²

With study in professional schools, including medicine, now within the resources of very large numbers of students, continued research leading to the development of increasingly effective testing instruments is essential. It would be desirable to have thorough studies made of such factors as: effects of coaching and cramming upon the medical aptitude test scores; relationships between test scores and ratings on interest inventories; relationship of test scores to drop-outs in the form of expectancy tables (that is, value of the tests in predicting survival in the professional school; not in predicting grades alone); correlations between test scores and interview ratings; role of personality traits in medical school performance and survival (that is, degree of emotional stability, degree and type of motivation, introversion-extroversion, dominance-submission, kinds and strengths of values, etc.). Admittedly, research on these and other personality traits would be difficult and long; but they might, nonetheless, prove significant.

TEST OF APTITUDE IN LAW

General Characteristics. Tests in this field are aptitude tests in the same sense as are those in the field of medicine. Psychologists concerned with the problem of testing aptitude for legal study agree that the following abilities are most important: reading rapidly and comprehending relatively difficult material, rapid memorizing and accurate recall, reasoning by analogy, discriminating between the relevant and the irrelevant in a mass of facts, reasoning inductively and deductively, facility in using and acquiring a vocabulary. Legal aptitude tests thus far constructed attempt to measure most or all of these abilities in some degree.

¹² Kandel, *op. cit.*, p. 20.

*The Iowa Legal Aptitude Test*¹³ (1948) includes the following seven parts: analogies, reasoning, opposites, relevancy, mixed relations, memory, and information. (1) The first of these requires the examinee to select from a group of six words the two which have to each other the same relationship as two key words. (2) The second part presents a series of stated premises as true; the examinee must decide whether a given conclusion is necessarily true, probably true, necessarily false, probably false, or indeterminate. (3) The third is the usual test of word opposites, with the addition that the subject must recognize specified parts of speech. (4) The fourth consists of a legal case which is presented for study; the student must determine whether each of the principles presented would be relevant to the argument of the plaintiff, the defendant, or to neither. (5) Part five requires the selection, from a group of six words, the three which are most closely related in some way. (6) Memory, part six, is tested by having the examinee study a detailed judicial opinion, prior to starting on part one; approximately two hours later, after completing part five, he is required to answer multiple-choice questions on significant facts, without referring back to the case. (7) Subtest seven consists of true-false statements concerning common facts of law which, it is assumed, students interested in the legal profession are apt to have acquired to a greater extent than others.

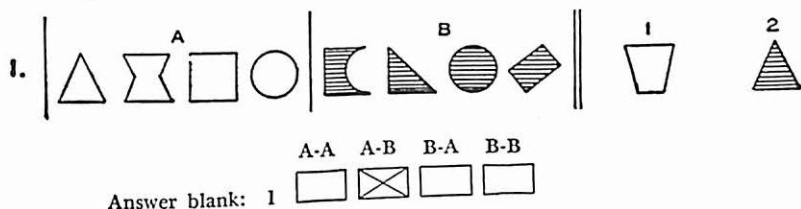
The *Law School Admissions Test*,¹⁴ introduced in 1948, is undergoing modification with experience. It is being used by the law schools of about forty-five universities. This test has six parts: (1) Principles and Cases; (2) Data Interpretation; (3) Reading Comprehension; (4) Debates; (5) Best Arguments; (6) Paragraph Reading (speeded reading, to detect the word in each paragraph that spoils that paragraph's meaning).

Although the parts of this instrument are named somewhat differently, they are designed to evaluate much the same abilities as the Iowa and others: understanding of words and word relationships, ability to read with understanding and analytically, capacity to reason in terms of verbal and nonverbal symbols. An unusual aspect of this test is the inclusion of nonverbal reasoning materials, as shown in Figure 13.3.

¹³ By M. Adams, L. K. Tunks, and D. B. Stuit, Bureau of Educational Research and Service, State University of Iowa, Iowa City.

¹⁴ Educational Testing Service, Princeton, N. J.

Directions: In each of the following problems you will be given two groups of figures, A and B, followed by two numbered figures. You must decide what characteristic it is that distinguishes the A figures from the B figures.



In the sample given above, you will note that all the figures under A are white; those under B are striped. The problem is to classify the figures under 1 and 2 as belonging to group A or to group B. The figure numbered 1 is white and therefore is an A; the figure numbered 2 is striped and therefore is a B. On the answer blank a cross has been made in the box beneath A-B, showing that in the last column there is an A figure followed by a B figure.

For each question determine in which group (A or B) each of the numbered figures belongs. Then, on your answer sheet, record your answer under the pair of letters which shows the order in which the numbered figures appear—(A-A if both figures are A; A-B if the first figure is A and the second B; B-A if the first is B and the second A; B-B if both are B).

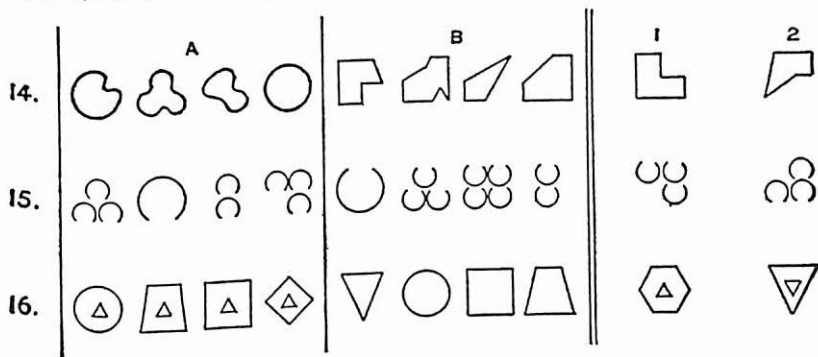


FIG. 13.3. Sample Items from the Law School Admissions Test. Educational Testing Service. (By permission.)

Evaluation of Legal Aptitude Tests. Prediction of scholastic achievement in law schools has been fairly widely studied.¹⁵ The two most valuable bases for prediction are marks in prelaw courses and per-

¹⁵ See H. R. Douglass, *University of Minnesota Studies in Predicting Scholastic Achievement*, Part II, Minneapolis: University of Minnesota Press, 1942; W. M. Adams, "Prediction of Scholastic Success in Colleges of Law," *Educational and Psychological Measurement*, Vol. 3, 1943, pp. 291-305, and Vol. 4, 1944, pp. 13-19. Also: A. B. Crawford and T. J. Gorham, "The Yale Legal Aptitude Test," *Yale Law Journal*, Vol. 49, 1940, pp. 1237-1249; W. B. Schrader and M. Olsen, "The Law School Admissions Test as a Predictor of Law School Grades," Princeton, N. J.: Educational Testing Service, 1950, p. 10.

formance on legal aptitude tests. Investigators are not in agreement as to which of these is the more significant, some having found that the former yields higher coefficients, while others have found the latter does so. In regard to either one of these predictive factors, the coefficients of correlation with first-year marks in law school were largely between .40 and .60. However, when prelaw college marks were combined with results of law aptitude tests, the multiple correlations with law-school marks were raised to the high .60's, and in some instances to the high .70's.

It may appear that many of the coefficients mentioned above are lower than should be expected of a valid aptitude test. There are several reasons why they are not higher. First, students *admitted* to many of the law schools are a selected group, the less competent, presumably, having been denied admission. The result is restriction in range of performance both on the aptitude tests and in law school courses. Thus, the averages of test scores and of law-school marks rise while their deviations (spread) decrease. This statistical situation results in lowering the correlation coefficient, the size of which is affected by the extent to which both sets of scores vary above and below their means. It follows, therefore, that the more selected a group is, the less well will the predictive factor (test scores) correlate with the criterion (law school grades).

A second factor is the disparities between academic and marking standards that prevail among colleges which give prelaw education. Consequently, college grades do not provide a uniform basis (variable) for correlation with the criterion, law-school grades. A common denominator is needed to supplement and check on undergraduate records if they are to be used as predictors.

A third factor contributing to depressed correlation coefficients is the variation in standards and marking among law schools themselves. There is no reason to suppose that uniformity or comparability is lacking only in undergraduate colleges.

The foregoing three factors are a strong argument for establishing *specific validity* rather than trying to determine *general validity*. That is, validity should be established for each law school individually, with regard to both prelaw grades and aptitude test scores.

Another important aspect of this problem is the fact that original selection of test content was a matter of face-validity. The materials for the tests were included because they seem appropriate to teachers

and practitioners of law, since the items are believed to require mental activities analogous to those used in the study and practice of the profession. This hypothesis, seemingly reasonably (and not unlike that adopted for other types of tests), should be subjected to investigation. Some of the questions to be answered are these: What is the predictive value (validity) of each of the several parts? What is the reliability of each part? How significant is the speed factor? What is the significance, in the tests, of technical vocabulary and legal information? Are the test batteries (requiring 3 to 4 hours) too long, so that fatigue and satiation influence the scores? To what extent do coaching and cramming affect test performance?

In spite of existing inadequacies, the available tests for selection of students of law are sufficiently far along in development and are helpful enough to warrant their use in conjunction with prelaw college grades. This conclusion is warranted by the significant validity coefficients found in some universities, by the high coefficients of multiple correlation, and by the ability of the tests to identify a very large portion of the candidates who show high promise and of those who show low promise.

TESTS OF APTITUDE FOR TEACHING

General Characteristics. It is extremely difficult to devise tests in this field because professional preparation of teachers is not so narrowly defined as is the case, for example, in law or medicine; and because teaching encompasses such a wide range of subject matter and educational levels. Probably the most useful single psychological test to predict quality and level of performance in the preparatory course for teaching is a test of general intelligence (or scholastic aptitude), since preparation for teaching is essentially academic. Equally important are the personality traits, essential in successful teaching, that may be discerned by means of interest inventories and personality tests, including projective techniques. As yet, however, no definitive psychological studies have been made of those traits.

The *Coxe-Orleans Prognosis Test of Teaching Ability* (1930) is intended to measure capacity to learn the subject matter taught in normal schools and in other institutions that educate teachers.¹⁶ The test consists of five parts: general information; student's observation of

¹⁶ World Book Co., Yonkers, N. Y.

classroom practices; brief lessons followed by questions in educational psychology, educational measurement, principles of teaching; comprehension of professional reading materials; problem situations, followed by questions to test the student's evaluation and response to each one.

The *National Teachers Examinations*¹⁷ (annual editions since 1940) are designed for use in the selection of teachers and for the appraisal of students in training. They are essentially subject-matter tests (including problems and applications) for the evaluation of relative adequacy of preparation of candidates for teaching positions in elementary and secondary schools. The battery of tests has two major divisions: the Common Examinations and the Optional Examinations.

Common examinations include the following tests, intended to measure professional knowledge, breadth of perspective on contemporary life, and reasoning ability.

Professional information: general principles and methods of teaching; educational psychology; child development; guidance and personnel services; evaluation of materials; philosophical and historical developments relating to contemporary education.

General culture: history, literature, and fine arts; science and mathematics; English expression.

Nonverbal reasoning: "abstract figures" are used, similar to those found in the matrices tests (discussed in an earlier chapter). The figures are incomplete; the examinee is required to discover the principle of arrangement.

Optional examinations, in addition to the foregoing, are offered in specialized areas, as follows: education in the elementary school (grades 1 through 8), early childhood education (through grade 3), biological sciences, English language and literature, industrial arts, mathematics, physical sciences, social studies, and physical education.

Evaluation of Teaching-Aptitude Tests. In one very important respect, these aptitude tests differ from those in law and medicine. Whereas the last two are administered *before* professional education is begun, teaching aptitude tests are administered *during* or *after* professional education to measure the professional knowledge and understanding the candidate has acquired.

¹⁷ Educational Testing Service, Princeton, N. J.

The Coxe-Orleans test, for which data are available, does what it sets out to do about as well as aptitude tests in other fields; for its authors have found correlation coefficients with scholastic grades in normal schools to vary from about .50 to about .85. The National examinations are also effective tests as measures of knowledge and understanding acquired and developed in the course of professional education.

Authors of these tests are careful to emphasize, however, that knowledge of subject matter and understanding of teaching procedures and of human behavior do not necessarily indicate one's ability to apply these effectively in the actual teaching process. Other aspects of the candidate's personality need evaluating if future classroom effectiveness is to be estimated: motives for teaching, emotional stability, social values, ability to communicate and establish contact with others (rapport), attitude toward and concept of the self.

It is quite improbable that any single personality scale will be able to yield useful results for all these traits. Some efforts, however, are being made to construct tests that will provide information regarding the non-intellective personality aspects essential in teaching. One of these is *The Case of Mickey Murphy: A Case-Study Instrument for Evaluating Teachers' Understanding of Child Growth and Development*.¹⁸ Another is the *Minnesota Teacher Attitude Inventory*, designed to predict how well the prospective teacher will get along with pupils.¹⁹ Both are pencil-and-paper tests that will have to be validated against competently rated classroom performance of many teachers who took the tests before beginning their professional careers.²⁰

TESTS OF SCIENCE AND ENGINEERING APTITUDES

General Characteristics. Scientific aptitude is not a special talent in the sense that musical aptitude, for example, is thought to be. Scientific aptitude may be characterized best as the application of general intellectual capacity to scientific materials. A test of scientific

¹⁸ By W. R. Baller. University of Nebraska Press, 1948.

¹⁹ By W. W. Cook, C. H. Leeds, and R. Callis. Psychological Corporation, 1951.

²⁰ It will have occurred to the reader that the tests of aptitude for teaching should, strictly, have been included and discussed under "proficiency tests" in the next chapter. They are included here, however, because they deal with one of the higher professions.

gebra), formulation of scientific relationships in algebraic terms, physical science information, arithmetical reasoning, scientific vocabulary, and comprehension of mechanical relationships and problems (presented in pictorial form). The authors present statistical details which show that their test has a reasonably high degree of validity when college marks in introductory engineering subjects are taken as criteria. The highest validity coefficients were found with marks in physics and chemistry, whereas the lowest were found with marks in "manufacturing processes" and drafting. This fact suggests that the instrument is more of a test of aptitude for the more purely scientific subjects than for applied engineering subjects of study.

Correlations with grade averages for each of the eight semesters ranged from .58 (first semester) to .26 (eighth semester), the decline in coefficients being steady from semester to semester, with one minor exception. To account for the decline from a coefficient of moderate

²¹ By D. L. Zyve, Stanford University Press, Stanford, California.

²² By B. V. Moore, C. J. Lapp, and C. H. Griffin, Psychological Corporation, New York.

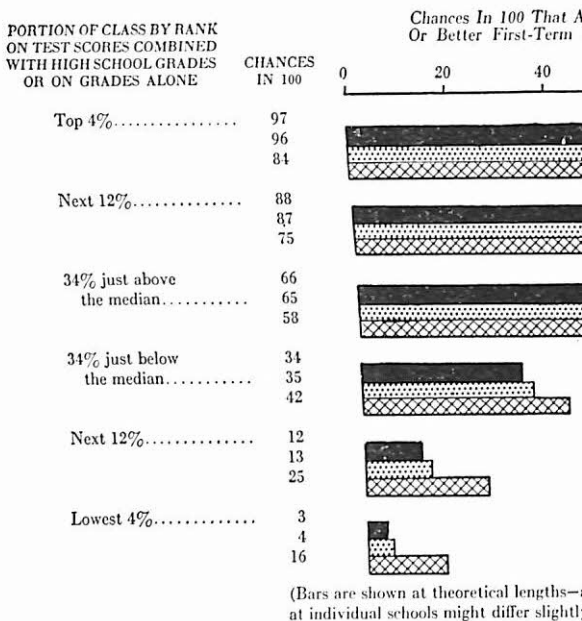


FIG. 13.4. Prediction of Scholastic Success in a C
Test Scores Combined with High-School Grad

* Data, courtesy of Dr. W. B. Schrader, for 721 enroll
Fall of 1948 at Carnegie Institute of Technology, Co
the University of Pennsylvania.

aptitude, therefore, should be regarded as a device intended to indicate the probability of success in scientific studies and occupations, without implying that such a test measures psychological functions which are essentially different from those used in other types of mental activity.

An early illustration of this type is the *Stanford Scientific Aptitude Test* (1930) which is intended for high-school seniors and college students.²¹ Its author states that the subtests evaluate experimental bent; clarity of definition; suspended *versus* snap judgment; reasoning; inconsistencies; fallacies; induction, deduction, and generalization; caution and thoroughness; discrimination of values in selecting and arranging experimental data; accuracy of interpretation; and accuracy of observation. Though this test has received more than passing attention and more than occasional use, it has not been subjected to a searching process of standardization and analysis necessary to establish its reliability and validity. Thus, unfortunately, while the subtests include items that are cleverly conceived and interesting, this device cannot be used in counseling with sufficient assurance of predictive value. At present, this instrument is principally of historical interest.

A more recent device is the *Engineering and Physical Science Aptitude Test* (1943-1951) compiled from a series of previously developed and standardized aptitude tests.

COLLEGE BOARD TESTS

Scholastic Aptitude Test

Mathematical Section (SAT-M)

Verbal Section (SAT-V)

Advanced Mathematics Test

Composite Score—an average of scores on the SAT-M and the Advanced Mathematics Test.

LEGEND:

Predictor

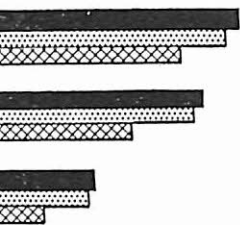
R or r**

 Composite (SAT-M and Advanced Math) combined with high school grades.....	.68
 SAT-M combined with high school grades.....	.66
 High School grades.....	.46

**Weighted average correlation coefficients against first-term engineering college grades

Student Will Earn Average
Engineering College Grades

60 80 100



actual lengths
(in inches.)

Group of Five Engineering Colleges by College Board
Tests. Educational Testing Service (By permission.) *

Engineering freshmen tested during their first week in the
Cornell University, Lehigh University, Rutgers University, and

magnitude to one that is very low, the usual influencing factors may be operative: increasingly homogeneous groups of students, reduction of individual differences due to earlier handicaps or advantages, increasing professional interest and motivation in students, unreliability of grading, etc. But here is an instance in which expectancy tables would provide more information than correlation coefficients, and probably would indicate that this aptitude test has greater predictive value than the correlations suggest.

The *Pre-Engineering Ability Test* (1952) has only two parts: comprehension of scientific materials and general mathematical ability. The first part requires reading and answering questions on scientific selections, tables, and graphs. The items in mathematics include arithmetic, algebra, and geometry (including analytic geometry). While it is maintained by the authors of this test that specific factual knowledge is not needed for the comprehension items, there is no doubt that familiarity with some of the materials, vocabulary, problems, and thinking in the physical sciences is an advantage.

Evaluation of Science and Engineering Aptitude Tests. The primary purpose of the instruments in this field of testing is to measure the most important elements of aptitude which are significant in the guidance and selection of pre-engineering students. Some of the current tests measure a fairly large variety of factors believed to be associated with success in the study of science and engineering. Others measure only a few. The tendency now is to reduce the number of subtests and to improve these, so that the smaller number have as much predictive value as the larger.

These aptitude tests, that correlate about .55 with first term grades, are more significantly correlated with tests of general ability (.70 and higher). And the latter themselves have as high predictive value as the former. The two types thus have much in common. Yet they are not perfectly correlated; nor will each type necessarily be as effective as the other in the case of any specific student. Better results, therefore, will be obtained if both types are administered, one to supplement the other, and if, in addition, the reasons are sought for significant discrepancies that might be found between the two scores of any candidate.

The desirability of investigating *specific* validity is indicated for these tests, just as in the case of law and medical school tests; for the

complex problems of intellectual and nonintellectual behavior are encountered in all areas when psychologists deal with learning, schooling, motivation, standards, etc.

Results obtained with current tests of engineering and scientific aptitude can be most helpful when analyzed and interpreted by a qualified professional counselor or psychologist. It is not the total score alone that might be useful. A profile representing scores on subtests can be even more valuable; but it is necessary to go still further in the analysis of an individual's performance. That is, a detailed analysis of performance on each type of subtest and item may reveal an individual's strengths and weaknesses that can then be evaluated in an interview with the subject, in the light of other available educational and psychological information regarding him. While numerical ratings obtained on some engineering and scientific aptitude tests will be of significance, considering their present validity coefficients, greater value from them will be derived in counseling by a qualified person who interprets the results in the light of his educational, psychological, and occupational insights. This is, admittedly, a subjective approach in part; but the development of this type of aptitude tests has not yet progressed far enough to warrant sole dependence upon their objective aspects.

INTEREST INVENTORIES

An individual's aptitudes and abilities ordinarily are not so highly specific that he can be given guidance solely on the basis of aptitude tests. Motivation, determined by one's interests, values, and preferences, may be the deciding factor in the selection of a course of study or of an occupation. Often, however, persons make their selections through chance influences rather than because of self-evaluation and knowledge of the field. Several devices, therefore, have been prepared to assist in the process of self-evaluation; and the results obtained with these are to be used in conjunction with data obtained from other sources. We shall briefly describe two widely used instruments, one of which scores an individual's interests in and preferences for a limited number of *areas*, whereas the other scores the individual's responses with reference to *specific* occupations.

The Kuder Inventories. The first of these is the *Kuder Preference Record* (1934-1951), in several forms, to be used with persons from

grade 9 and onward.²³ The inventory includes 168 items, each of which lists three activities; for example:

Build bird houses
Write articles about birds
Draw sketches of birds

The subject indicates which he likes most and which least. The items, covering a wide range of activities, are scored to yield a profile representing ten areas: namely, mechanical, computational, scientific, persuasive, artistic, literary, musical, social service, clerical, and outdoor. A person's scores on each of the ten areas is converted into a percentile rank; and the resulting profile is analyzed with a view to determining in which, if any, areas the individual's interests and preferences are strong.

Identification of areas does not, of course, determine the one or several specific occupations a given person should enter on the basis of preference. The author of this Preference Record, therefore, has listed under each area of preference numerous specific occupations which are associated with it and which should receive special consideration. Since it is not uncommon for an individual's profile to show more than one area of strong preference, Kuder has enumerated occupations of possible interest under various pairs of preferences such as mechanical-artistic, mechanical-scientific, scientific-artistic, scientific-social service, persuasive-literary, etc. Some of the listings are based upon actual data; many, however, are based upon the author's judgment of consistency between occupational duties and activities included in the inventory.

The Strong Inventories. *The Strong Vocational Interest Blank* (1927-1951),²⁴ available in separate forms for men and for women from age seventeen onward, contains 400 items dealing with likes and dislikes in occupations, school subjects, amusements, activities, and peculiar personality traits; with order of preference of activities, importance of factors affecting one's work, order of preference of men

²³ Science Research Associates, Chicago. Kuder has also developed a *personal Preference Record* (1948-1953). This inventory consists of five scales describing different types of personal and social activities. These scales are regarded as being relatively independent of the vocational inventory. The two inventories are intended to supplement each other in vocational counseling and employee placement.

²⁴ Stanford University Press, 1938.

TABLE 43
Percentile Ranks of Mean Scores in Various Occupations²⁵
(Kuder Preference Record)

Occupation	Outdoor	Mech.	Com.	Sci.	Pers.
Men					
Accountants	42	27	94	43	53
Civil Engineers	65	57	71	66	18
Lawyers and Judges	41	19	48	36	57
District Forest Rangers	84	50	56	54	32
Office Managers and Chief Clerks	37	29	77	39	57
Sales Managers	32	30	32	38	87
Women					
Librarians	41	46	30	35	44
Physicians	84	77	47	86	22
Social and Welfare Workers	51	43	31	43	54
Occupational Therapists	70	88	32	58	26
Trained Nurses	51	57	43	69	33
Secretaries	51	43	49	41	64
Sales Clerks					
Occupation	Art.	Lit.	Mus.	Soc.	Cler.
Men					
Accountants	34	63	56	36	83
Civil Engineers	67	55	42	25	55
Lawyers and Judges	43	82	61	48	61
District Forest Rangers	55	54	36	53	29
Office Managers and Chief Clerks	45	65	55	43	66
Sales Managers	41	57	59	47	43
Women					
Librarians	66	83	52	29	38
Physicians	60	58	45	63	11
Social and Welfare Workers	59	70	46	81	25
Occupational Therapists	81	45	61	58	14
Trained Nurses	58	46	54	75	26
Secretaries	50	64	61	39	58
Sales Clerks					

(or women) one would like most and least to have been, positions one would like most and least to hold in an organization, comparison of interests between paired items, and self-rating of present abilities and traits.

The purpose of the inventory is to find the extent to which an individual's interests and preferences agree with those of successful persons in specified occupations: forty-one for men, twenty-five for women. The inventory, scored separately for each occupation, yields

²⁵ From Examiner Manual, Science Research Associates, Chicago, 1951, pp. 14 ff. (By permission.)

Group Key number Form M

VOCATIONAL INTEREST BLANK FOR MEN (Revised)

By EDWARD K. STRONG, JR.
Professor of Psychology, Stanford University
Published by STANFORD UNIVERSITY PRESS, Stanford University, California

Part I. Occupations. Indicate after each occupation listed below whether you would like that kind of work or not. Disregard considerations of salary, social standing, future advancement, etc. Consider only whether or not you would like to do what is involved in the occupation. You are not asked if you would take up the occupation permanently, but merely whether or not you would enjoy that kind of work, regardless of any necessary skills, abilities, or training which you may or may not possess.

Draw a circle around L if you like that kind of work.
Draw a circle around I if you are indifferent to that kind of work.
Draw a circle around D if you dislike that kind of work.

Work rapidly. Your first impressions are desired here. Answer all the items. Many of the seemingly trivial and irrelevant items are very useful in diagnosing your real attitude.

1 Actor (not movie)	L	I	D	46 Jeweler	L	I	D
2 Advertiser	L	I	D	47 Judge	L	I	D
3 Architect	L	I	D	48 Labor Arbitrator	L	I	D
4 Army Officer	L	I	D	49 Laboratory Technician	L	I	D
5 Artist	L	I	D	50 Landscape Gardener	L	I	D

Part II. School Subjects. Indicate as in Part I your interest when in school.

101 Algebra	L	I	D	106 Repairing a clock	L	I	D
102 Agriculture	L	I	D	107 Adjusting a carburetor	L	I	D
103 Arithmetic	L	I	D	108 Repairing electrical wiring	L	I	D
104 Art	L	I	D	109 Cabinetmaking	L	I	D
105 Bookkeeping	L	I	D	110 Operating machinery	L	I	D

Part III. Amusements. Indicate in the same manner as in Part I whether you like the following or not. If in doubt, consider your most frequent attitude. Work rapidly. Do not think over various possibilities. Record your first impression.

137 Golf	L	I	D	234 Progressive people	L	I	D
138 Fishing	L	I	D	235 Conservative people	L	I	D
139 Hunting	L	I	D	236 Energetic people	L	I	D
140 Tennis	L	I	D	237 Absent-minded people	L	I	D
				238 People who borrow things	L	I	D
				239 Quick-tempered people	L	I	D
				240 Optimists	L	I	D

Part IV. Activities. Indicate your interests as in Part I.

Part V. Peculiarities of People. Record your first impression. Do not think of various possibilities or of exceptional cases. "Let yourself go" and record the feeling that comes to mind as you read the item.

Part VI. Order of Preference of Activities. Indicate which three of the following ten activities you would enjoy most by checking (✓) opposite them in column 1; also indicate which three you would enjoy least by checking opposite them in column 3. Check the remaining four activities in column 2.

	1	2	3	
201 () () ()				Develop the theory of operation of a new machine, e.g., auto
202 () () ()				Operate (manipulate) the new machine
203 () () ()				Discover an improvement in the design of the machine
204 () () ()				Determine the cost of operation of the machine
205 () () ()				Supervise the manufacture of the machine

Part VII. Comparison of Interest between Two Items. Indicate your choice of the following pairs by checking (✓) in the first space if you prefer the item to the left, in the second space if you like both equally well, and in the third space if you prefer the item to the right. Assume other things are equal except the two items to be compared.

Work rapidly.

321 Street-car motorman	()	()	()	Street-car conductor
322 Policeman	()	()	()	Fireman (lights fire)
323 Chauffeur	()	()	()	Chef
324 Head waiter	()	()	()	Lighthouse tender
325 House to house canvassing	()	()	()	Retail selling

Part VIII. Rating of Present Abilities and Characteristics. Indicate below what kind of a person you are right now and what you have done. Check in the first column ("Yes") if the item really describes you, in the third column ("No") if the item does not describe you, and in the second column (?) if you are not sure. (Be frank in pointing out your weak points, for selection of a vocation must be made in terms of them as well as your strong points.)

	YES	?	NO
301 Usually start activities of my group	()	()	()
302 Usually drive myself steadily (do not work by fits and starts)	()	()	()
303 Win friends easily	()	()	()
304 Usually get other people to do what I want done	()	()	()
305 Usually live up the group on a dull day	()	()	()

TABLE 44. Sample items from Strong Vocational Interest Blank.
Standard University Press. (By permission.)

ratings which are considered to be indicative of the subjects' interest in each. It is possible, also, to score the inventory for six "occupational groups," rather than for specific occupations, in instances where one wishes to know which broad fields of occupation are indicated. The inventory may, in addition, be scored for "non-occupational" interests, found to be useful in guidance: namely, interest maturity; masculinity-femininity (based upon percentile scores of males and females on interest items); occupational level (differences in interests between "laboring men," on the one hand, and "business and professional men earning \$2500 and upwards a year," on the other); studiousness ("factors which contribute to scholastic achievement that are not measured by intelligence tests").

First Ten Items on Vocational Interest Blank	Percentage of "Men-in- General" Tested			Percentage of Engineers Tested			Differences in Percentage between Engineers and Men-in-General			Scoring Weights for Engineering Interest		
	L	I	D	L	I	D	L	I	D	L	I	D
Actor (not movie)	21	32	47	9	31	60	-12	-1	+13	-1	0	1
Advertiser.....	33	38	29	14	37	49	-19	-1	+20	-2	0	2
Architect.....	37	40	23	58	32	10	+21	-8	-13	2	-1	-1
Army officer.....	22	29	49	31	33	36	+9	+4	-13	1	0	-1
Artist.....	24	40	36	28	39	33	+4	-1	-3	0	0	0
Astronomer.....	26	44	30	38	44	18	+12	0	-12	1	0	-1
Athletic director	26	41	33	15	51	34	-11	+10	+1	-1	1	0
Auctioneer.....	8	27	65	1	16	83	-7	-11	+18	-1	-1	2
Author of novel..	32	38	30	22	44	34	-10	+6	+4	-1	1	0
Author of technical book.....	31	41	28	59	32	9	+28	-9	-19	3	-1	-2

TABLE 45. Determination of weights for an Occupational Interest Scale: Engineering²⁷

The procedure followed in determining the significance of and the weights to be given particular responses is, in outline, as follows:²⁶

1. Blanks are filled out by an adequate sampling of an occupation and by a much larger sampling of "men-in-general."
2. Responses to each item are tallied according to the three categories: like, indifferent, dislike.
3. Frequency of responses to each item, in each of the three categories.

²⁶ Cf. E. K. Strong, *Vocational Interests of Men and Women*, Stanford University Press, 1943, p. 74. The formula for determining weights, p. 611.

²⁷ From E. K. Strong, *Vocational Interest of Men and Women*. Stanford University Press, 1943, p. 75. (By permission.)

gories, are calculated in terms of percentages. Table 45 presents the percentages for the first ten items on the blank, as responded to by "men-in-general" and by engineers.

4. Weights are calculated for each of the frequencies of each of the three responses—like, indifferent, dislike—to each item. Table 45

First Ten Items on the Vocational Interest Blank	Scoring Weights for Engineering Interest			Responses of an Engineer to the Ten Items			Scores for Engineering Interest Obtained by This Engineer	Scores Obtained by This Engineer on Interest Scales for				
								Lawyer	Life Insurance Salesman	Minister	Y.M.C.A. Secretary	Accountant
Actor (not movie)	-1	0	1	..	..	X	1	-1	-1	-2	-1	-1
Advertiser.....	-2	0	2	..	..	X	2	1	-1	0	-2	-1
Architect.....	2	-1	-1	..	..	X	-1	1	1	0	0	0
Army officer.....	1	0	-1	..	X	..	0	0	0	-1	0	0
Artist.....	0	0	0	..	X	..	0	0	-1	1	1	0
Astronomer.....	1	0	-1	X	..	..	1	0	0	2	0	0
Athletic director	-1	1	0	..	X	..	1	0	0	1	0	0
Auctioneer.....	-1	-1	2	..	X	..	-1	-1	0	0	0	0
Author of novel.	-1	1	0	..	X	..	1	0	0	0	0	0
Author of technical book.....	3	-1	-2	X	..	..	3	0	-1	-1	-1	1
Total 10 items.....							+7	0	-3	0	-3	-1
Total 400 items.....							+182	+23	-115	-91	-134	-33
Standard score.....							67	36	10	11	3	16
Rating.....							A	B	C	C	C	C

TABLE 46. Scores obtained by an engineer for engineering interest; also scores for interest in five other occupations, illustrating the method of scoring the Strong Interest Blank.²⁸

shows that 21 percent of "men-in-general" but only 9 percent of engineers would like to be actors. For this item, a calculated weight of 1 is obtained; and it is given a minus sign because fewer engineers marked "like" for this item. Essentially, the weight given an item depends upon the difference in percentages found between the score of an unselected sampling of persons and that of the occupation for which the item weight is being determined. Table 46 illustrates scoring and weights for the ten items in Table 45.

Evaluation of Interest Inventories. The Kuder and the Strong inventories are not tests of aptitude; they indicate only the extent of simi-

²⁸ From E. K. Strong, *op. cit.* p. 75. (By permission.)

larity between interests and preferences expressed by those being examined and those of persons successfully engaged in the specified occupations or areas. It has been found that persons successfully engaged in a particular occupation or group of occupations have, in general, some communality of preferences and dislikes which differentiate them from persons in other occupations and groups. The principle underlying these and similar inventories is, then, that an individual who has a pattern of preferences and dislikes similar to the distinguishing aspects of a given group or occupation has a greater chance of finding that type of activity congenial and, hence, of succeeding in it, provided, of course, that he has also the degree of aptitude required. The differentiating value of the Strong inventory is illustrated in Figure 13.5, which contrasts rather markedly divergent occupational groups in respect to interests.

These two inventories differ in their approach to the problem. The Kuder attempts to identify important aspects of vocational interests in broad occupational areas. A probable occupation or a restricted list of occupations may be suggested by a score in a single area. Or still another restricted list may be suggested by a combination of two or more areas. The Strong, on the other hand, aims primarily to provide patterns of preferences that distinguish one specific occupation from others, as rated by the same inventory.

Reliability and Validity. While reliabilities, in terms of internal consistency, of both the Kuder and Strong inventories are satisfactory (the coefficients falling generally between .85 and .95), the more difficult problem, by far, is demonstration of their validity. For the Strong inventories, a number of different criteria were used: namely, mean scores and standard deviations of criterion (specific occupational) groups compared with a general sample, relationships of scores to other test scores (of a variety of types), correlations with grades in schools and colleges, completion of occupational training, ratings of success in work, earnings (in sales work), persistence in occupations, job satisfaction, and differences between occupational groups. Excepting the correlations found with tests of intelligence, educational achievement, personality traits, and with school and college grades, these criteria, all already more or less familiar to the student, have been found to be related significantly to the scores on the Strong inventories.

The Kuder inventories have used much the same types of validation

Aptitude Tests: Fine Arts and Professions

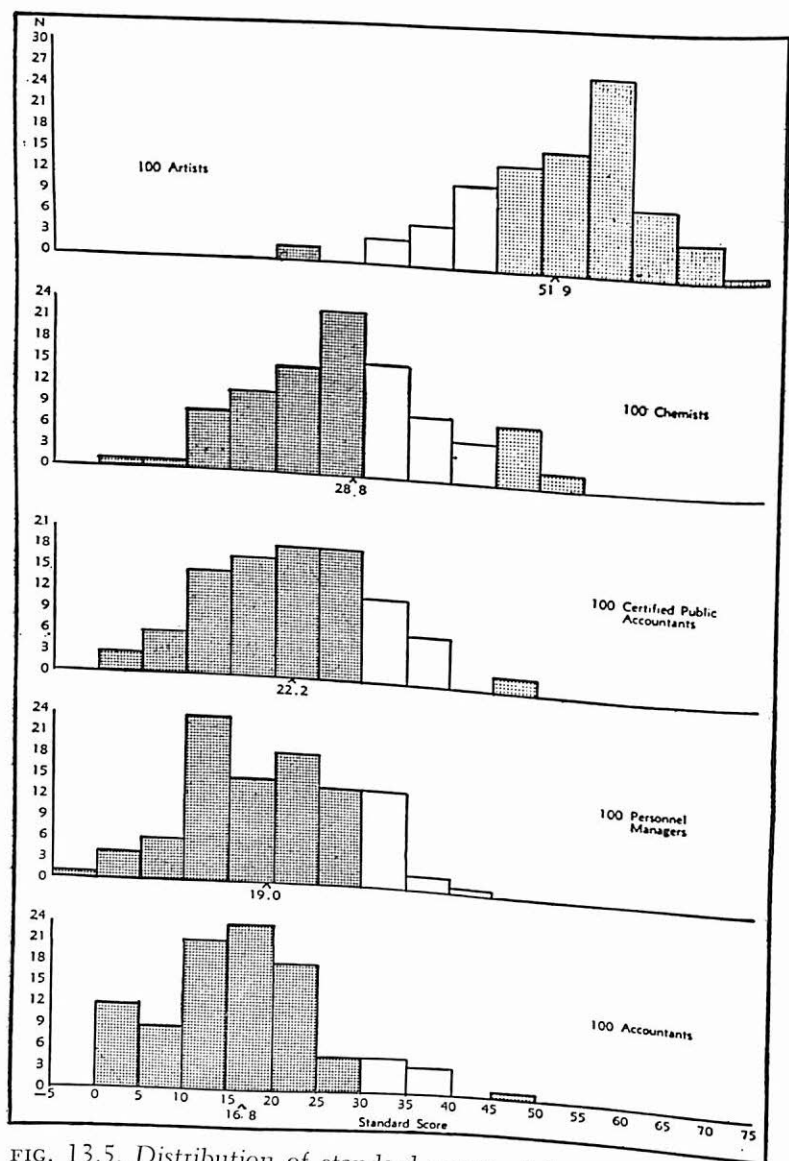


FIG. 13.5. Distribution of standard scores of five occupations on Artist Scale. From E. K. Strong, *op. cit.*, p. 111. (By permission.)

criteria. In addition, scores in the separate areas have been intercorrelated. These reported coefficients have ranged from $-.34$ (scientific with persuasive) to $+.50$ (clerical with computational). Ideally, of course, low intercorrelations are desired; for then the scores will have greater differentiating value.

Both inventories were validated, also, by means of item analysis. Essentially, the problem would be to examine an adequate number of responses to each item by known groups, to determine which of them are selected and rejected most often by persons strongly interested or successful in a given area (Kuder) or occupation (Strong). For example: do high scorers in the "Scientific" area select an item more often than the low scorers in the same area? Do high scorers in a given area tend to choose the same items? How frequently, say, do successful engineers select a particular item?

In this connection, Kuder has found that "mean profiles" for occupational groups ". . . indicate in general that the names assigned to the various scales are appropriate in terms of the type of occupation entered as well as in terms of the activities for which the scale is scored. Chemists are found to be particularly high on the scientific scale, writers on the literary scale . . .," etc. He reports, also, that scores on the preference record are consistent with students' choices of occupations and of college curricula. Strong likewise finds that his inventory differentiates persons successfully engaged in one occupation from those successfully engaged in others. A second criterion employed by Strong is the extent to which the inventory distinguishes between persons who are successful in an occupation and those not so successful in the same work. In this respect, too, Strong reports satisfactory results.

The scores obtained with the Kuder preference record have in general a low correlation with educational achievement tests, with course marks, and with tests of general intelligence. The Strong inventory, when correlated with tests of intelligence, yielded coefficients which were low or negligible, some of them even being negative. These results indicate that there is but a very weak relationship, or none at all, between general ability and occupational interests. When, therefore, these and similar devices are used in guidance, they must be supplemented by measures of general ability, among others, especially in attempting to determine *occupational level*—for example, as between

clerk and accountant, or machinist and mechanical engineer, technician and scientist.

Stability of Interests. A very important question asked in the early years of these and similar inventories, and still asked by some critics today, is the extent to which occupational interests change with the inseparable factors of age and experience, especially after adolescence; hence, the question of the dependability and predictive value of these instruments. Strong and others have published studies comparing interests of several age groups of boys and men, from age 15 to 55. It was found that there was only moderate agreement between interests of 15-year-olds and 25-year-olds, and somewhat closer correspondence between 15-year-olds and men above age 25. This interesting phenomenon may be due to shifting interests, searching, trial-and-error, and unrealistic concepts of the self, accompanied by physiological changes, between the ages of 15 and 25. After age 25, it appears that men on the whole settle into more realistic and stable interests and activities. More important than age differences were the persistent occupational differences throughout the entire age range. The major weakness of these studies is that they did not follow the same individuals over a period of years to determine how many shifted significantly in their interests.

In several other investigations, however, the same persons were examined twice: one group as college freshmen and again after a period of nine years; another group as college seniors and again after ten years. Correlations between the two sets of scores for the first group averaged .56, while the average coefficient for the second was .71. Both of these indexes are quite significant; yet they indicate that there are many appreciable changes in interests and preferences during undergraduate years, while there is a much stronger tendency toward stabilization after college graduation.

Data dealing with stability of interests and preferences of a group of high-school pupils of both sexes (in grades 10 and 11), who were re-examined as college seniors, yielded a correlation coefficient of .52—again only a moderate degree of correspondence.

Studies of the Kuder, dealing with the same question, indicate that over periods of about a year, the scores on this inventory do not change greatly, in the case of an adult group. This is a finding to be expected. Other longitudinal studies indicate that only moderate changes in score occur during the high-school and college years. This

is not entirely in accord with the data on the Strong; but the difference is probably due to the fact that the latter measures interests for specific occupations, whereas the Kuder does so for broad areas.

It is not surprising that changes in scores are found over a period of time. Many considerations and factors other than ability and preferences influence vocational choices which, in turn, have their influence upon subsequent interests. These influencing factors are often subtle, unpredictable, and frequently unknown. Under these circumstances, the results and information provided by these inventories are very creditable.

Applicability to Different Age Groups. With whom should such occupational and interest inventories be used? Obviously, they can have validity only for persons whose lives have been long enough and varied enough to have provided them with experiences of the kind which will enable them to choose between the alternatives presented by each item in the inventories. The Kuder Preference Record, concerned with broad areas of interest, is standardized for high-school students (beginning with grade 9), college students, and adults at large. The Strong Vocational Interest Blank, being concerned primarily with specific occupations, is intended for ages seventeen and over. Since the Strong inventory has been based upon responses of adult men and women, more valid and useful results will be obtained with adults than with persons in their teens. Since the Kuder record has been standardized with high-school and college students, as well as with adults, it may be used appropriately with adolescents. But even so, the interests, values, and attitudes of adolescents are still in a state of flux and are as yet not fully developed; hence the results of the preference record, when used for guidance purposes, must be interpreted with this factor in mind.

With either a high-school or a college student, the scores and the profile obtained with the sounder inventories in this category are useful as an introduction to the study of occupations which involve activities of the sort for which he has indicated a preference, and during interview and counseling to check the individual's choice of an occupation against his expressed interests and preferences. For purposes of guidance at the secondary-school and college levels, it appears that the Kuder has the greater value because it is less specific.

Both instruments under discussion are intended to provide measures of motivation in various fields of study and work. Their applica-

some would have succeeded. Such a selective procedure is directed toward the improvement of the percentage of successful selections. Admissions officers of colleges and professional schools also can employ the same principle if they are satisfied with a mechanical and impersonal procedure. For this purpose, tests of even relatively low validity (e.g., coefficients of .30 or .40) have some value when supplemented by expectancy tables.

In this connection, the "selection ratio" has been used.³¹ The term is defined as the ratio of the number of persons selected to the number tested. Thus, if 200 individuals were examined for a specified type of employment or for admission to a school of law, for example, and if only the highest 100 were to be selected, the selection ratio would be .50. Whenever there is a reliably established positive relationship between the test scores and the criterion of performance, any degree of selectivity will to some extent yield a group of persons whose promise is greater than that of the whole group, the degree of superiority depending upon the size of the test's validity coefficient. For example, assume we have an aptitude test that correlates .50 with the criterion (e.g., level of performance in a professional school). Assume also that of the candidates for selection, only one half will be chosen. The selection ratio, therefore, is .50. We might ask ourselves the following question, among others. What are the chances that the scholastic performance of those in the selected group will place them in the upper half of the levels of performance of an unselected group? In the present instance the chances are 67 in 100. Any validity coefficient of a test and any selection ratio may be thus used in combination to predict the probable *average* performance of a selected group at any level of success.³²

Otherwise stated, assume that only the upper 50 percent of an unselected group are to be chosen. What percent of those retained are likely to prove satisfactory if the validity coefficient is of a certain size? For example, if the coefficient is .20, the percent satisfactory will probably be 56; if the coefficient is .30, the percent is 60; if the co-

³¹ See H. C. Taylor and J. T. Russell, "The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection: Discussion and Tables," *Journal of Applied Psychology*, Vol. 23, 1939, pp. 565-578.

³² For details of this method see Taylor and Russell, *op. cit.* Also, M. Smith, "Cautions Concerning the Use of the Taylor-Russell Tables in Employee Selection," *Journal of Applied Psychology*, Vol. 32, 1948, pp. 595-600.

efficient is .40, the percent is 63; if the coefficient is .70, the percent is 75. Tables of relationships have been calculated for combinations of selection ratios ranging from .05 to .95 and coefficients of validity from .00 to 1.00. These tables provide a useful general index for estimating the probable value of a test of given validity in a given situation.

We must re-emphasize that this technique yields indexes applicable to a group, but it does not show what will necessarily happen in the case of a particular individual within that group. The utilization of the selection ratio does, however, add to the utility of aptitude tests when group trends and a minimum of elimination from a job or course of study are the first concerns. When the individual himself is of first concern—as he is to clinician and counselor—aptitude-test results must be used only as one kind of information in a total picture.

STEPS IN AN APTITUDE TESTING PROGRAM

Although tests for only very few of the many occupations, including professions, have been presented, the procedures and problems are the same in all fields. While tests are available in certain specified areas (e.g., mechanical, scientific, verbal, etc.) the particular combinations and emphases for a given vocation will depend upon the characteristics and demands of each occupation. And certain occupations will need especially devised tests in addition. For example, the study of nursing, dentistry, and medicine have much in common, but each has its own aspects. Thus, many occupational aptitude tests require “custom made” batteries that consist of some parts common to others and some parts specific to themselves.

When instituting a program of aptitude testing for the prediction of performance and selection of individuals for any area, however, the procedures are the same as for any other. The following steps are indicated, if one is starting at the beginning:

- (1) Analysis of the task to determine the traits that contribute to success or failure.
- (2) Development or choice of tests and other evidence for trial purposes.
- (3) Use of tests, etc., with an experimental group of subjects.
- (4) Selection of validity criteria and obtaining validity data regarding performance of the experimental group:

- (a) Simple correlations of each test and each of the other ratings with performance ratings or scores;
 - (b) Multiple correlations of the combined scores and ratings with performance ratings or scores;
 - (c) Percents of successful and unsuccessful individuals getting high scores and low scores on the tests;
 - (d) Critical scores ("cutting scores") which show the test score or rating level at which the probable percentage of failures in actual performance begins to be so high that persons falling below it are poor risks.
- (5) Selection of tests and procedures based upon validity and other standardization data.

TESTS OF EDUCATIONAL ACHIEVEMENT

SCOPE

A test of educational achievement is one that is designed to measure knowledge, understandings, or skills in a specified subject, or group of subjects, taught in school. The test might be restricted to a single subject, such as arithmetic or history; or it might be a "battery," that is, tests of several areas of subject matter yielding a separate score for each subject and a total score for the several subjects combined.

Tests of educational achievement differ from those of intelligence in that: (1) the former are concerned with the quantity and quality of learning attained in a subject of study, or group of subjects, after a period of instruction; (2) the latter are general in scope and are intended for the measurement and analysis of psychological *processes*, although they must of necessity employ some content which has been acquired and which resembles some of the content found in achievement tests.

Most tests of achievement are devoted very largely to the measurement of the amount of information recalled or skills and techniques acquired. In more recent years, however, an increasing number have been devised to measure such educational results as problem solving, drawing inferences from subject matter (inductive thinking), applying generalizations to specific situations and problems (deductive thinking), attitudes and appreciations developed by the study of course materials, and practices and skills developed in the study of a

of standardized tests often disclose such a situation. The psychologist may wish, also, to utilize achievement-test results in evaluating a pupil's school progress as part of a total case history. In all these situations, the comparative ratings provided by a standardized achievement test are more valuable than are the subjective marks of individual teachers, though the latter are not to be disregarded.

DERIVED INDEXES

Scoring of educational tests does not involve unique problems or principles. Their raw scores may be converted into the familiar percentile ranks and standard scores; or they may be converted into "educational age" or "achievement age," by means of a table of age-norms. Raw scores may be converted, also, into "grade equivalents," using a table of grade norms.

The reader is already familiar with the concept of the norm and with the method whereby it is derived. Thus, if a pupil's raw score on a standardized test of the four fundamental arithmetical processes gives him a grade equivalent of 4.5, it means that his achievement on this test is equal to that of the average achievement of pupils who have completed half of the fourth grade. By means of this index, obtained for each of the several subjects of instruction, it is possible to get a profile for each pupil and thereby to evaluate his general level of achievement, the evenness or unevenness of his performance, his weakness and strength in the type of learning measured.

The educational age (EA) designates a pupil's average standing in the several subjects in which he has been tested. For example, assume that a pupil has been given a test battery on which his scores were as follows: reading rate and comprehension, 9-year level; arithmetic fundamentals, 10-year level; spelling, 9-year level, language usage, 10-year level. His EA would be the average of these, or 9.5. Each of the separate ages represents the measured average level of educational achievement of the pupils of that age. The EA is a composite, representing a pupil's average achievement.

Achievement age (AA) designates a pupil's level of performance in a single school subject. The "subject age" is sometimes used synonymously with AA. Thus, if a pupil gets a rating of achievement age 10 on an arithmetic test, his level is equal to the norm of 10-year-old pupils, as measured by the given test. If "subject age" is used, it would be said that this pupil's "arithmetic age" is 10. At present, it is a mat-

ter of personal choice whether to designate a pupil's performance on an arithmetic test as "achievement age (arithmetic)" or "arithmetic age."

Grade norms, educational ages and achievement ages derived for different tests are not always comparable, nor are they always applicable in a given school or community, since the standardization populations of these tests vary in respect to adequacy and representativeness. This is an important problem in achievement testing, since quality of education is far from uniform among various parts of this country. Rather than deriving only national norms, it is much more meaningful to present, in addition, separate norms for different sections of the country, and even for different types of communities (according to population). In connection with norms, it is necessary to recall the earlier distinction made between norms and standards (Chapter 2). Norms, in this instance, will represent the levels of manifested achievement in the subject matter being tested. They do not necessarily represent the level of learning that might be desirable or optimal—that is, a standard of performance.

It is necessary to have a quotient to accompany an educational or an achievement age. Hence, there are two types of quotients used with tests of educational achievement when EA or AA is found: namely, the educational quotient (EQ) and the achievement quotient (AQ). The latter is sometimes called the accomplishment quotient.

The educational quotient is the ratio of educational age to chronological age (EA/CA) multiplied by 100 to remove the decimal; that is, the individual's average level of measured learning in relation to what is expected on the basis of his life age. Theoretically, the "normal" EQ is 100; deviations above or below represent, respectively, superior or inferior school learning as compared with the individual's age group.

The achievement quotient is the ratio of educational age to mental age (EA/MA), multiplied by 100; that is, the individual's average level of measured learning in relation to what is expected on the basis of his mental level.¹ Due to marked individual differences in mental

¹ The AQ is also given as EQ/IQ. This is the equivalent to EA/MA, but in another form. Thus: $IQ = MA/CA$; $EQ = EA/CA$. Substituting for EQ/IQ, we have

$$\frac{EA/CA}{MA/CA}, \text{ or } EA/MA.$$

ability in any group of persons of the same life age, the mental age is regarded as a more reliable index of a person's learning capacity than is chronological age; therefore, the AQ is a more valuable index than the EQ in judging whether or not a pupil's school achievement is commensurate with the quantity and quality of learning that might reasonably be expected of him.

The educational quotient and the achievement quotient have been used not only when the composite EA is derived and taken as the numerator of the ratio, but also when only the age level for a single subject (subject age or achievement age) is used as the numerator. When this is the case, then the quotients should be read as "educational quotient in" (naming the school subject tested; e.g., arithmetic); "achievement quotient in"

The value and wisdom of using the AQ have been seriously questioned, for the three following reasons. (1) Since this index is derived from two separate tests, each of which has a certain degree of error or unreliability, it has a lower reliability than an index based upon either test alone. (2) Since the norms of the two tests, one of educational achievement and the other of general mental ability, have in all probability been derived from different standardization groups, they will not be strictly comparable, nor will their measures of dispersion. For example, unless the distributions of scores of the two standardization groups are approximately equal, a given EA (say, 10) and EQ (say, 110) will not be comparable with an MA of 10 and an IQ of 110. Unless the two distributions are equivalent, or nearly so, the two sets of apparently identical indexes will have different values, more or less. Hence, an index derived from them will not have the same meaning as either one. (3) Since courses of study in most schools are geared to pupils of average ability, bright pupils will not have an opportunity to learn up to a level consistent with their capacities; hence, they will be penalized and will get AQ's of less than 100. Slow pupils, on the other hand, will be working "over their heads," so to speak, and will more often get AQ's above 100.² In spite of these criticisms, however, and assuming that one is aware of its limitations, there are occasions when the AQ serves a useful purpose, especially in studying the educational problems of superior children.

² This criticism does not seem to have much merit; for what it says, in effect, is that the AQ helps to reveal an undesirable educational situation, so far as superior pupils are concerned.

TYPES OF ITEMS

Test items may be classified into one or another of the following major types: (1) simple recall; (2) two alternatives, such as true-false, right-wrong, yes-no; (3) multiple-choice; (4) completion; (5) matching; (6) analogies; and (7) check lists. The following examples illustrate these types.

- (1) *Simple recall:*
 What are the dates of World War I?
 What are the two main gases found in water?
- (2) *Two alternatives:*
 NaCl is the chemical symbol for common salt. T F
 Abraham Lincoln served two complete terms as president
 of the United States. T F
- (3) *Multiple-choice:*
 Scrooge is a character in
 1. Oliver Twist
 2. David Copperfield
 3. A Christmas Carol.
 An example of an appointed official is a
 1. congressman
 2. senator
 3. federal judge.
- (4) *Completion:*
 The executive head of the United States government is the
 _____, while the federal legislative bodies
 are the _____ and the _____.
- (5) *Matching:*
 Directions: After each name write the number of the topic
 which is intimately associated with that person.

(1) conditioned reflex	Titchener
(2) age scale for testing intelligence	Hall
(3) reaction-time experiments	Pavlov
(4) psychoanalysis	Cattell, J. M.
(5) psychology of adolescence	Freud
(6) existential psychology	Binet
(7) factorial analysis	
- (6) *Analogies:*
 Executive functions : President :: legislative functions :
 Hydrogen : H :: sodium :
- (7) *Check lists:*
 Directions: Place a check mark in front of each of the follow-
 ing items which is part of an automobile:

throttle
rudder
gear shift
periscope

generator
distributor
aileron
stabilizer

While these are the principal types of items used in tests of educational achievement, they are not found with equal frequency; true-false, multiple-choice and completion are the most common. There are also variations on some types. For example, tests of paragraph meaning (e.g., in literature, social sciences, physical sciences, and others) are quite common. The testee reads a paragraph and is then required to answer questions intended to show the extent to which he has comprehended the paragraph. The questions usually are in the form of true-false, multiple-choice, or completion. In arithmetic and other aspects of mathematics, the items in a standardized test usually are in the same form as those devised by the individual teacher: that is, the student simply provides the correct answer. In tests of some aspects of English, special types of items have been developed, as, for example, in tests of punctuation and capitalization. In the following illustration the necessary capitals and marks of punctuation are to be supplied.

why did you come home so early mary asked

THREE REPRESENTATIVE BATTERIES

The number and range of tests of educational achievement are very extensive; and they vary markedly in merit. Whenever a school or clinical psychologist is faced with the necessity of evaluating educational achievement of a group or of an individual, the selection of tests must be based upon their appropriateness to the problem at hand and upon their adequacy of standardization: range of grades covered, aspects and comprehensiveness of subject matter covered, reliability, validity, and norms. It is our purpose in this section to indicate the kinds of subtests found in only several of the more adequately constructed and standardized scales; for it is neither possible within a single chapter nor necessary to present a comprehensive list to familiarize the reader with the nature of these instruments.

The Stanford Achievement Tests.³ These provide batteries at four levels: primary, elementary, intermediate, and advanced. The primary

battery (for use at the end of grade 1, in grade 2, and in the first half of grade 3) includes tests of paragraph meaning, word meaning, spelling, arithmetic reasoning (concepts and problems), and arithmetic computation (the four fundamental processes).

The elementary battery (for grades 3 and 4) includes tests of paragraph meaning, word meaning, spelling, language (mechanics and usage), arithmetic reasoning, and arithmetic computation.

The intermediate battery (grades 5 and 6) and the advanced battery (grades 7, 8, and 9) include the same tests: namely, paragraph meaning, word meaning, spelling, language, arithmetic reasoning, arithmetic computation, social studies, science, and study skills.

The California Achievement Tests.⁴ These are designed for a very wide range of educational levels: grades 1 through 14. The four batteries (primary, elementary, intermediate, and advanced) all include the same general tests, with changing content and of increasing difficulty, of course. These tests are: reading vocabulary, reading comprehension, arithmetic reasoning (mathematics reasoning in the advanced battery), arithmetic fundamentals (mathematics fundamentals in the advanced battery), mechanics of English and grammar, and spelling. Each of these, with the exception of spelling, is divided into several subtests that are regarded as the component parts of the broader category of school learning.

The Iowa Tests of Educational Achievement. These are less comprehensive in scope than either of the foregoing batteries, being devised for mid-8th grade to mid-13th. There are nine subtests: understanding of basic social concepts; general background in the natural sciences; correctness and appropriateness of expression; ability in quantitative thinking; ability to interpret reading materials in the social studies and natural sciences, and literary materials; general vocabulary; and uses of sources of information.

Evaluation of the Stanford, California, and Iowa Batteries. These three batteries, briefly described, are among the most widely used.

³ By T. L. Kelley, R. Madden, E. F. Gardner, L. M. Terman, and G. M. Ruch. Published by World Book Co., Yonkers, N. Y., 1953.

⁴ By E. W. Tiegs and W. W. Clark. Published by California Test Bureau, Los Angeles, 1933-1951.

They, and other similar batteries,⁵ have much in common in regard to objectives, standardization numbers and range of population samples, reliability, and conception of validation.

These batteries have been current long enough so that their normative data, for each form, are based upon many thousands of cases: Stanford, 70,000 to 106,000; California 50,000 to 100,000; Iowa, 50,000. In each instance, the cases were distributed over a wide geographic area; particularly the Stanford, the norms of which were derived from school systems in 38 states.

Reliabilities of the component parts of all three are very largely within the range considered satisfactory:

California: .83 to .96

Iowa: .81 to .94

Stanford: .73 to .95

Validity of these and similar tests is "curricular" validity (content validity). This criterion of selecting test content is justifiable in this area. The ultimate value of an achievement battery will depend upon: (1) the thoroughness with which test materials were selected originally; (2) adequacy of item analysis to determine the discriminating power of each; and (3) the differentiating value of the norms, as between chronological age groups and school grade levels.

If an achievement test is valid, it measures what is actually being taught in the schools for which it is intended. The procedure followed in constructing the sounder achievement tests, therefore, is essentially as follows. The most widely used textbooks in the subject are thoroughly analyzed; a wide variety of course syllabi are analyzed; educational objectives of the courses of instruction are defined; an analysis is made of research findings relating to pupils' experiences, forms and levels of concepts, and vocabularies at successive ages and school grades. On the basis of these analyses, outlines of test content are prepared, in which proportions are allotted to all the skills, types of information, and understandings that are to be examined. The outlines of content and proportions are then submitted to subject-matter specialists for their judgment. After all this has been done, the first version of the test is ready for a series of try-outs, involving the usual standardization techniques, during which items are revised, added, and

⁵ See, as the most comprehensive single source, O. K. Buros (editor), *Third and Fourth Mental Measurement Yearbook*, 1949 and 1953, respectively.

eliminated, until satisfactory reliability and validity are achieved. To a considerable extent, these rigorous standardization procedures were used in constructing the three batteries under discussion, particularly so in the case of the Stanford.

These batteries are intended to be analytical of pupil achievement in these ways: (1) they measure continuity of educational growth over a wide range of school grades; (2) they reveal group or classroom differences in subject-matter, skills, or insights being tested; (3) they reveal differences in competence within a single individual; (4) they identify those pupils who are so markedly deficient in any area as to require more intensive testing, observation by their teachers, and diagnosis of the elements of deficiency, with a view to remedial instruction. The comprehensive survey batteries under discussion are not designed for this kind of deficiency diagnosis, though they contribute to this end in a limited degree. For example, if a pupil earns a very low score on the arithmetic test in the battery (for reasons other than poor general intelligence), a diagnostic test in arithmetic (either standardized or especially devised) should reveal in which of the four fundamental processes and at which levels he is deficient; in which number combinations or specific or isolated skills and understandings he is weak.

On the whole, achievement batteries of the quality of these three can contribute much to the study and solution of the types of educational and psychological problems presented under the section on "Purposes" in this chapter.

READING TESTS

Reading is the school subject to which most research has been devoted and for which the largest number of tests have been devised. This is quite understandable, since reading ability is the keystone of the usual school curricula through elementary and high schools, and also in most courses of study at higher levels. The numerous tests in this field have much in common, but they are not identical as to content, some being more comprehensive than others. Reading tests may be divided into three groups: achievement, diagnostic, and readiness.

Achievement. These tests are concerned with the measurement of how well the testee is able to read: that is, how rapidly and with what degree of understanding. For these two purposes, most current tests

measure rate of reading, sentence or paragraph comprehension, word meaning (vocabulary), and recall of details. Less often, reading tests measure knowledge of abbreviations, selection of key words, and use of an index or directory. The tests in this category are intended to measure achievement in each of the several areas and thereby to indicate, at the same time, the causes of poor reading on the part of individuals of whom a higher level of reading competency might be expected.

Diagnostic. Such tests are intended, through analysis, to isolate in detail the factors responsible for an individual's reading disabilities; that is, factors other than lack of general intelligence. The following tests and the elements in reading, analyzed by each, are representative of the approach to the problem and of the factors being isolated.

*Diagnostic Reading Tests.*⁶ Grades 7-13.

1. vocabulary: in the fields of English, mathematics, science and social studies.
2. comprehension: silent and auditory
3. rate of reading: general, social studies, science
4. word attack: oral and silent—identification of sounds, syllabication.

*Ingraham-Clark Diagnostic Reading Tests.*⁷ Grades 1-8.

1. word form and meanings: word likenesses and differences, auditory stimuli, visual stimuli, opposites, similarities, association.
2. sentences and paragraphs: following directions, answering questions, directly stated facts, qualified statements, inferences and conclusions, relevant and irrelevant statements, true and false deductions, selecting and classifying information, form and mechanics of organization and sequence of events.

*Durrell Analysis of Reading Difficulty.*⁸ Grades 1-6.

- | | |
|-------------------------------|-----------------------|
| 1. oral reading comprehension | 5. word pronunciation |
| 2. oral reading recall | 6. spelling |
| 3. silent reading | 7. handwriting |
| 4. word recognition | |

⁶ Committee on Diagnostic Reading Tests, 419 W. 119th Street, New York 27, N. Y., 1947-1952.

⁷ California Test Bureau, Los Angeles, 1929.

⁸ World Book Co., 1933-1937.

*Monroe Diagnostic Reading Examination.*⁹ Grades 1-5.

- | | |
|---|-----------------------------------|
| 1. alphabet repeating and reading | 5. mirror reading |
| 2. word recognition (reading) | 6. mirror writing |
| 3. letter reversals and inversions | 7. number reversals |
| 4. recognition of word and letter reversals | 8. word discrimination (auditory) |
| | 9. sounding words by syllables |

*Van Wagenen-Dvorak Diagnostic Examination of Reading Abilities.*¹⁰ Grades 4-16.

- | | |
|--|-----------------------------------|
| 1. rate of comprehension | 6. grasping central thought |
| 2. perception of relationships (analogies) | 7. retention of detail |
| 3. vocabulary in context | 8. integration of dispersed ideas |
| 4. vocabulary (not in context) | 9. drawing inferences |
| 5. general information | 10. interpreting contents |

Inspection of the preceding lists shows that although the several instruments differ in some respects in their content, they have much in common, with the exception of the Monroe which is concerned entirely with the sensory aspects and the mechanics of reading and writing, particularly with visual anomalies, and auditory and visual deficiencies that interfere with learning and with progress in reading. The Monroe test, however, is usually preceded and supplemented by results obtained with some of the usual tests of reading achievement. Emphasis in the diagnostic tests changes, of course, as the educational level becomes higher, the changes depending upon the skills and competence usually expected in the progressively higher school grades.

Readiness. Tests of *reading readiness* serve a still different purpose from either of the other two types, already described. Children are customarily admitted to the first grade on the basis of age, the assumption having been that when a child reached a specified age he was ready to begin the study of prescribed first grade subjects. Extensive investigation of individual differences has revealed, however, that an appreciable percentage of children are not ready to receive formal instruction in reading at the usually prescribed time due to immaturity

⁹ C. H. Stoelting Co., 1930.

¹⁰ Educational Test Bureau, 1939-1940.

of some specific types of perceptions. Reading readiness tests, therefore, have been devised in an effort to identify children who are not yet mature enough to receive and benefit from instruction in this fundamental subject.

The tests described below are representative of the more widely used instruments that have been found valuable in determining reading readiness.

*Gates Reading Readiness Tests.*¹¹ Grade 1.

1. following directions in marking pictures (ability to listen, understand directions, and remember)
2. word matching (identification of visual word patterns)
3. word perception (selecting one word from among four)
4. rhyming (auditory perception)
5. naming letters and numbers

*Harrison-Stroud Reading Readiness Tests.*¹² Kindergarten and Grade 1.

1. making visual discriminations
2. using context
3. making auditory discriminations
4. using auditory clues in identifying items
5. using symbols

*Metropolitan Readiness Tests.*¹³ Kindergarten and Grade 1.

1. word knowledge (selecting pictures that correspond to words)
2. understanding of and response to oral directions (selecting pictures in response to sentence-long directions, involving sustained attention)
3. information (same as 2, but more elaborate, involving more vocabulary and names of common objects)
4. matching (visual perception, involving selection of pairs of identical pictures of common objects)
5. knowledge of numbers
6. copying (simple geometric forms and less complex numerals and capital letters)

¹¹ Bureau of Publications, Teachers College, Columbia University, 1939.
¹² Houghton Mifflin Co., 1950.
¹³ World Book Co., 1933-1950.

*Monroe Reading Aptitude Tests.*¹⁴ Ages 5-6 to 8-11.

- | | |
|---|---|
| 1. memory of orientation of forms | 10. word articulation |
| 2. ocular-motor control and attention | 11. speech facility (speed or repetition of a word or phrase) |
| 3. visual memory | 12. vocabulary by word association (naming animals, foods, toys) |
| 4. motor speed | |
| 5. motor steadiness | 13. sentence length and use of language (telling a story about a picture) |
| 6. auditory discrimination of words | 14. motor test of handedness (writing name with hand) |
| 7. sound blending | 15. laterality of hands, feet, and eyes |
| 8. vocabulary (recognition of pictured objects) | |
| 9. auditory memory | |

Tests of reading readiness are based upon observation and analysis of aptitudes and abilities that have been found to be involved in learning to read by school beginners. Inspection of each of the subdivisions of the listed tests shows that there is extensive agreement among them regarding what is to be measured, in spite of somewhat different names and emphases. Of the four instruments listed, the Monroe is most detailed analytically; it places much more emphasis upon visual, auditory, and motor functions than do others. On the whole, in broad terms, reading readiness is measured in terms of sensory development and acuity, language development and interest, curiosity about the environment (information, vocabulary, numbers), and, to a lesser extent, motor control and speed (since learning to write is collateral with learning to read).

These and comparable tests have demonstrated their value. The percentile ratings on the Gates test, for example, indicate whether a child may be expected to experience much, little, or moderate difficulty in learning to read. The Metropolitan provides scores and ratings (A to E) that indicate probabilities of successfully learning to read. The tests are significantly correlated with later achievement in reading in the early stages. The Monroe yields, for example, an exceptionally high coefficient (.75) with reading achievement at the end of the first grade.

¹⁴ Houghton Mifflin Co., 1935.

his problems aloud, while the examiner (usually the teacher) records the pupil's responses, step by step. Thus the errors, faulty reasoning, and incorrect methods may be detected.

Much more work remains to be done in this field of objective testing before it will have reached the level of development of tests in other areas, such as intelligence, some aptitudes, and reading. Perhaps a major reason for the lag in the field of arithmetic is the wide variation among schools in content, method, and emphasis in teaching this subject; hence, the non-applicability of a test standardized for general use.

TESTS AT HIGH-SCHOOL AND COLLEGE LEVELS

At the secondary school level, tests are available in nearly all subjects of study. They vary considerably in quality, with regard to both adequacy of content and technical standardization. Among the more satisfactory of these are the comprehensive series published by the Cooperative Test Division of the Educational Test Service,¹⁵ and the Iowa Placement Examinations.¹⁶ They are tests, very largely, of information and fundamental skills; but in some instances, as in English, they measure comprehension and appreciation.

The College Entrance Examination Board has constructed achievement tests to be used as one criterion for college admission. Tests are available in social studies, foreign languages, physical and biological sciences, and mathematics. Some of these instruments differ from others in that they not only lay emphasis upon dealing with subject matter learned in high-school courses, but they test to some extent the student's ability to apply basic principles in new situations. Thus, they are intended to serve as measures of what has been learned and as predictors of future performance with similar course materials in college.

Since 1937, objective tests of educational achievement for college graduates have been in process of standardization and are being used under the auspices of the Graduate Record Office.¹⁷ These constitute a comprehensive series in the academic fields designed to show the level and extent of a student's general academic knowledge in comparison with groups having had a similar education. The items are

¹⁵ Princeton, N. J.

¹⁶ Bureau of Educational Research and Service, University of Iowa.

¹⁷ Educational Testing Service.

constructed for the purpose of examining the student's knowledge, ability to solve problems in general fields of study, and ability to exercise judgment based upon knowledge of academic material. The general tests cover seven subject-matter fields, plus vocabulary and effectiveness of expression, these being: mathematics, physics, chemistry, biology, social studies, literature, fine arts. There are, in addition, specialized examinations to test students in their subjects of major study. The general and specialized tests have two major purposes; they may be used to evaluate the student's progress during his undergraduate career, and they may be used as criteria for admission to graduate study. For the latter purpose, the specialized examination is regarded as being of particular value.

TESTS OF APTITUDE IN SPECIFIC ACADEMIC SUBJECTS¹⁸

In the fields of mathematics and foreign languages, at the high-school level, a few tests have been constructed which are intended to predict pupils' achievement in these subjects. They are aptitude tests in the sense that they use restricted and specialized types of materials in an effort to predict performance in a very limited field. Tests of aptitude in specific school subjects are constructed on the well-recognized principle that a sample of performance in a given subject, obtained under controlled conditions, provides significant evidence of prospects of learning that subject. These aptitude tests are not to be confused with educational achievement tests. The former are prognosis tests, intended for those who have not studied the subject; the latter are measures of learning after a period of formal instruction.

In *algebra*, several tests are available.¹⁹ They include subtests such as the following: arithmetic problems; analogies; number series; use of formulas; brief lessons, of a relatively simple level, dealing with principles and problems of the kind actually encountered in the first course in algebra. Thus the subtests include materials some of which were taught in earlier courses of study, while others are *similar to*

¹⁸ Although these are classified as aptitude tests, they are being included in this chapter because they are intimately associated with achievement in specific courses of study at the secondary school level.

¹⁹ *Iowa Algebra Aptitude Test*, by H. A. Greene and A. H. Piper, Bureau of Educational Research and Service, State University of Iowa, Iowa City; *Orleans Algebra Prognosis Test, Revised Edition*, Yonkers, N. Y.: World Book Co.

those found in the study of algebra; and still others are actual *learning samples* of the subject-matter.

There are several prognosis tests also in *plane geometry*, constructed on the same principles as those in algebra.²⁰ They include parts on: reading geometry content, algebraic computation, algebraic and arithmetical reasoning, visualization, and brief lessons in geometry followed by test items.

Aptitude tests in *foreign languages* likewise are constructed on the principle that samplings of actual learning provide the best criterion for predictive purposes.²¹ They consist of short lessons dealing with grammatical aspects, vocabulary exercises, and translations. The tests, it appears, are devised to predict achievement in the grammar-translation type of foreign language course.

Evaluation of High School Tests. The authors of aptitude tests in these subjects, and other investigators, report validity coefficients that range, approximately, from .60 to .80, although lower coefficients have been reported on occasion. Nevertheless, in spite of some of the relatively high coefficients, some specialists in the teaching of foreign languages maintain that the available tests are useful only for courses of the traditional grammar-translation type, and chiefly for the purpose of homogeneous grouping. It is maintained, also, that pupils who do poorly on these tests might, however, benefit from a "functional" course in foreign language. This is not the place to evaluate the teaching of foreign language; but in evaluating a test in this field, it is necessary to know the kind of course of study with which the test is concerned.

The principal criticisms of the aptitude tests in algebra and geometry are that some do not show sufficiently high reliability, and most place too much weight on rate of work, thus adversely affecting their predictive value. If a prognosis test is to predict potential power or level of achievement, the speed factor should be eliminated; or available testing techniques should be utilized to obtain separate

²⁰ *Iowa Plane Geometry Aptitude Test*, by H. A. Greene and H. W. Bruce, State University of Iowa, Iowa City; *Orleans Geometry Prognosis Test, Revised Edition*, Yonkers, N. Y.: World Book Co.

²¹ *Foreign Language Prognosis Test*, by P. M. Symonds, Bureau of Publications, Teachers College, Columbia University, New York; *Luria-Orleans Modern Language Prognosis Test*, Yonkers, N. Y.: World Book Co.

speeded and unspeeded scores, even though the correlation between speed and power is very high (but not perfect, nor nearly so).

TESTS OF MORE COMPLEX EDUCATIONAL OBJECTIVES

For the greatest part, the items and content of the educational tests thus far discussed are useful primarily, though not solely, for testing recall of information or use of certain relatively fixed skills. More recently, however, some authors have devised tests which are intended to measure performance and educational outcomes on a higher level: tests of evaluation, interpretation, critical thinking. An example of this type is Wrightstone's *Test of Critical Thinking in the Social Studies*²² which includes the following three types: (1) obtaining facts; (2) drawing conclusions; (3) applying general facts. The first part requires the student to select certain items from a group of facts in accordance with specific directions. The correct answer is given to each question, the task of the pupil being to match answers and given facts. In this part of the test, the pupil is required to obtain facts from graphs, tables, maps, indexes of a textbook type, and to locate information in books, magazines and newspapers in the ways required in a library. This is, in short, a test of ability to utilize a variety of kinds of materials in order to acquire information. Part two, called "Drawing Conclusions from Facts," is a reading test which requires the pupil to evaluate a number of conclusions drawn from given data. All necessary information is provided, so that recall is not involved. The procedure consists of reading a paragraph, then matching several given statements with it to determine which of them is appropriately drawn from the paragraph. Part three, "Applying General Facts," is much the same as part two, for the pupil is required to generalize and apply facts presented in a paragraph to be read.²³

The Wrightstone test is open to criticism regarding its content and the degree to which it achieves its stated purpose; but it is a type of measure worth noting because it is a departure from the much more common instruments which seek to measure only information. It

²² By J. W. Wrightstone. Published by Bureau of Publications, Teachers College, Columbia University. 1939.

²³ The directions in the test itself will help clarify this part: "Below each paragraph are two sets of statements about the paragraph. In the left-hand column are five statements. Three of these statements will help you to understand the references in the right-hand column. Select a statement from the left-hand column which best explains a reference in the right-hand column."

an achievement test, educators, guidance counselors, and psychologists must closely scrutinize its manual for evidence of sound standardization procedures, statistical (score) validity, and the bases of content validity.

Tests of educational achievement are particularly valuable in the primary and elementary grades where they are used to measure pupils' basic skills chiefly in reading, spelling, and arithmetic, and in extent of vocabulary. For these purposes standardized tests, with their norms of performance and their diagnostic means, provide teachers and others with superior instruments for the measurement of pupils' progress in universal fundamentals.

In other respects and at higher educational levels, standardized tests have been used to excess in this country. Many of them—some circulated by reputable publishers—are very poor devices even from the standpoint of accepted standardization procedures alone. Most others—even those that meet the demands of standardization procedures—are of limited value because they test very little, if anything, beyond discrete items of information and routine skills (e.g. arithmetical processes, punctuation). If these tests, therefore, are used as the sole or principal measure of educational outcomes, they fail to evaluate what should be the most important and the basic aims of education. The mere acquisition of facts, which are soon forgotten, is not the main purpose of education.

Modern aims emphasize such objectives as the development of concepts and attitudes, critical thinking, analysis and synthesis of materials, creativity, originality, problem solving and the like. With relatively few exceptions, current tests of educational achievement do not attempt to measure these abilities; and the few that do make the attempt are by their very nature limited in the extent to which they can test the attainment of these educational objectives. For these latter tests (e.g., the Graduate Record, the Wrightstone, and the Watson-Glaser) present a number of problems or situations for each of which the facts are given, as are the possible solutions, inferences, logical explanations, evaluations, or interpretations; and it is the examinee's task only to *select* certain of these as being correct, or sounder or more relevant than others. There is nothing creative, original, or spontaneous in such a task; individual critical thinking is reduced, often to a minimum. At best they are tests of ability to discriminate between arguments and to recognize assumptions. While such tests, if well con-

ceived, are greatly superior to the simple measurement of information, their limitations must be recognized.

On the whole those achievement tests which assist the teacher, the psychologist, or the school counselor in discerning sources of an individual's educational difficulties will be most useful also for clinical and guidance purposes. Most others, above the primary and elementary grades, may be useful in obtaining an index of progress in school subjects, chiefly with respect to amount of information or skills acquired.

TESTS OF PROFICIENCY

Tests in this category deal with achievement level in occupational areas rather than in subjects of school study. They are, in fact, specialized achievement tests. For this reason they are designated differently and separately classified.

Proficiency tests are available especially in bookkeeping, shorthand, typewriting, and other subjects in business education. A battery of proficiency tests may, however, be built from existing instruments that were originally devised for another purpose. Such batteries have been labelled "custom built." For example, analysis of a particular type of office job may show that proficiency in four areas is necessary: namely, arithmetical computation, spelling, perceptual speed, and memory span (immediate recall). As a first step, it might be possible, in this instance, to utilize in combination four existing tests already standardized. But, of course, specific validation studies for this particular situation would be necessary. On the other hand, the nature of the job might be such as to require the development of tests adapted to this specific situation.

Another type of proficiency test is the *work sample*. This is nothing more nor less than testing the candidate by having him actually do a piece of the work required by the job to be filled. Examples of this would be a replica of a telephone switchboard, an airplane pilot trainer (on the ground), a machine, typewriting and shorthand. To have any merit beyond a randomly selected task chosen by each employer, these work sample tests must be carefully selected so as to be representative of the job, and scoring must be placed on a basis that is specifically defined and minimizes subjectivity.

A third type is the test of *analogous functions*. Instead of asking the candidate to produce a sample, he is subjected to tests which measure

functions that are analogous to those that are necessary in the job. For example, if manual speed and precision are required, speed of tapping, peg boards, and stylus tracing might be used as tests.

Generally, the available proficiency tests have not been standardized in the same sense as tests of intelligence and some others have been. The reasons for this state of affairs are, presumably, that many proficiency tests are devised for limited and specific situations only, jobs even under the same name (e.g., stenography) vary considerably as to requirements, and it is difficult to obtain large enough and representative enough population samples. Yet, it should be possible to develop these devices to a higher level of objectivity and validity. A few of the tests in this category have been so developed.

One of these is the *Seashore-Bennett Stenographic Proficiency Test*,²⁶ which is to be used for selecting, training, and up-grading employees. This is a *work sample* test, consisting of five letters that are dictated at three different rates. The testee then transcribes the letters in a form that might be mailed out. The final product is scored for: (1) neatness and cleanness of typing; (2) arrangement of the letter; (3) quality of stroke; (4) typing errors and erasures; (5) errors in English; (6) changes from the original in wording and meaning. The letters to be dictated and transcribed are on records, thus keeping the quality and rate of dictation constant for all testees. Samples of superior and poor transcriptions are provided as guides for scoring.

Reliability of the Seashore-Bennett test ranged from .80 to .91 (correlation of scores on letters 1, 3, 5 with scores on 2 and 4). Also the scores of two clerks on the individual letters of 52 stenographers yielded coefficients of .74 to .97. Validity was estimated in terms of supervisors' ratings of 52 stenographers, the coefficient of correlation being .49.

On the whole, the merit of proficiency tests is that, in spite of lack of normative data and adequate validation for widespread use, they have been more carefully and analytically prepared than the usual tests devised by individual instructors, personnel officers, and employers. It will be found, too, that *specific* rather than general validity of proficiency tests will prove to be most valuable.

²⁶ Psychological Corporation, 1946.

INTELLIGENCE TESTS AS CLINICAL INSTRUMENTS

CLINICAL psychology involves the intensive psychological study of individuals in which testing, observation, interview, and history taking are utilized, in whole or in part. The purpose of such study, in some instances, is to determine the causes of each individual's mal-functioning and to prescribe suitable educational and psychological measures to deal with the problem. These measures may include educational changes and adaptations, manipulation of the person's environment (in one or more of several ways), vocational guidance, or psychotherapy.

Not all cases coming to a psychological clinic, however, are problems of maladjustment. They may be individuals who want whatever objective and psychologically valid information might be obtainable regarding their general and specific abilities, their interests, and sometimes, their personality traits.

All types of tests have become indispensable instruments in the broad practice of clinical psychology. Every test can be clinical in a literal sense, since it helps to analyze an individual's abilities, to obtain a more nearly complete description of his strength and weakness. In this chapter we shall deal, however, only with two of the most widely used tests of intelligence—the Stanford-Binet and the Bellevue—and with several other instruments devised particularly for the determination of mental abnormality or deterioration. In a later chapter we shall present some tests of personality and their clinical uses. While tests of specific aptitudes and educational achieve-

ment are often essential in the study of some individuals, these do not present the clinical problems and possibilities found in intelligence and personality tests; the clinical uses of aptitude and achievement tests, especially in cases of educational and vocational counseling, are much more obvious and direct. Nor shall we deal further with tests for infants and preschool children. These are not used in so wide a variety of clinical problems as are the Stanford-Binet and the Bellevue; their organization, furthermore, is such that they readily lend themselves to analysis of performance in terms of sensory, motor, perceptual, language, and social development.

FACTORS WHICH AFFECT TEST PERFORMANCE

Before discussing any specific tests, it is necessary to point out the factors which can affect an individual's performance on the occasion of any psychological examination and of which the clinical examiner must be aware. These factors—or sources of error—are of two general kinds: *intrinsic* and *extrinsic*. By the former, we mean those factors which are *within the subject himself*; by the latter, we mean those *outside the subject*. Lack of IQ constancy, in some instances, must be evaluated in the light of possible disturbing factors, discussed below.

Intrinsic Factors. Such factors that affect test performance include: (1) organic difficulties such as: defective hearing or vision; disability or enervation due to malnutrition, localized or generalized infections; glandular dysfunctions; acute or chronic illnesses which lower an individual's level of performance; (2) emotional factors as evidenced in: lack of interest, lack of seriousness, deliberate deception, negativism; inhibition due to shyness or lack of confidence; hyperactivity and restlessness; neuroses and the more severe forms of mental disturbances; (3) language handicaps; (4) speech defects.

The presence or absence of organic difficulties may be inferred from the subject's history or from reports provided by observers, such as teachers; but their existence and degree can be adequately determined only through physical examination. During the process of testing, however, the experienced psychological examiner is often led to suspect or infer the presence of organic difficulties. The report of the psychological examination will, in such instances, include a description of the subject's behavior and performance which gave

rise to the suspicion or inference. This description must be taken into account in the interpretation of the test results.

As for emotional factors, the psychological examiner must be able to discern their presence in the subject's behavior, or, where the nature of the instrument permits, in his actual performance on various parts of the test itself. This latter aspect will be discussed specifically in connection with the several tests presented subsequently in this chapter. Discernment of negativism, deception, shyness, lack of interest, and lack of seriousness is a form of subtle clinical insight which can be developed only through experience in the actual testing of a variety of individuals who manifest these traits in varying degrees, and of others who do not manifest them at all. Such contrasting individuals provide the necessary basis of comparative evaluations.

Extrinsic Factors. Factors of this kind affecting test ratings include the following: (1) accidental factors, such as errors in time limits, broken pencils, distracting noises; (2) scale errors inherent in the tests themselves due to imperfect standardization; (3) scoring errors due to judgment of the examiner or to the marginal character of the response—that is, a response which is on the borderline between the acceptable and nonacceptable; (4) skill of the examiner who must not only be thoroughly familiar with the instrument he is using, but who must be able to establish the rapport necessary to elicit the subject's best performance. These sources of error signify that the examiner must be highly qualified and must know the standardization and limitations of his instrument.

The recognition of factors that can affect test performance is always essential for sound interpretation of test results. It is especially necessary that these factors be given due weight when the psychological examiner is confronted with the problem of making a diagnosis of mental deficiency. Such a diagnosis should not be made on the basis of an intelligence test alone; for attempts to use only IQ's in diagnosing mental deficiency not only disregard the necessity of having additional information about the individual, but they also reveal a failure to recognize the probability of "errors of measurement" due to intrinsic and extrinsic factors.¹

¹ Some states, for example, unwisely have written into law the requirement that a child must have a certain IQ (say, 75) or lower to be admitted into a special class for mentally retarded children.

Tests provide a situation in which the subject is psychologically observed as well as being scored in accordance with objective standards. In addition to observation and standardized scoring, analysis of an individual's performance on subtests and specific items can often yield significant information concerning his mental status and mode of functioning. The following sections on the Stanford-Binet and the Bellevue scales will be devoted to this aspect of test interpretation.

THE STANFORD-BINET SCALE

Uses. In schools, guidance centers, and clinics, the problems for which intelligence tests have been and are being widely used are the following: diagnosis of mental deficiency and mental retardation; determination of mental levels of delinquents; diagnosis of intellectual disturbance and deterioration; differential diagnosis (that is, mental traits and profiles of various clinical groups); evaluation of the intelligence of maladjusted children; evaluation of the intelligence of children with special disabilities in learning; determination of mental superiority; educational guidance for children and others who are not "problem" cases, as that term is ordinarily used; vocational guidance, for which the determination of intelligence level is one essential.

Functional Analysis. For these and other clinical problems it is not sufficient to find only the mental age and the intelligence quotient, since an individual's performances on the several parts of an intelligence test are not entirely uniform. He will be above his general average on some and below on others. The MA and IQ are composites, whereas in the analysis of an individual case we must know, in addition, the components that have yielded these indexes. The reader should recall that tests are standardized on the basis of performances which are characteristic of representative *groups* of persons at particular age levels, and it rarely happens that a given individual conforms to his age group in all respects. Two identical mental ages may differ functionally in terms of their components, as inspection of the results on two Stanford-Binet tests will readily reveal. Two children, for example, each with a mental age of ten, may vary in respect to basal year, extent of scatter (see below), and items passed or failed at each of the several levels.

Functional analysis of Stanford-Binet test results is not readily apparent because the items are grouped according to age levels rather

than psychological processes.² It is possible, however, to draw certain conclusions, regarding an individual's strength and weakness, through analysis of the test items. For example, levels may be determined in the following functions: visual perception of form (three-hole form board, copying a square and a diamond); visual imagery (copying a bead chain from memory, and paper-cutting test); visual memory (reproducing geometric designs); thinking (absurd sentences and arithmetical problems); memory span and attention (recall of digits and of meaningful sentences); ability with abstractions (verbal similarities and differences); reasoning (plan of search, picture absurdities, and picture interpretation); word knowledge (vocabulary); concept formation (definitions of abstract words). Analysis of successes and failures in these or similar terms will assist the examiner in determining the psychological components of average or superior mental-age rating, or the deficiencies responsible for an inferior rating.

Analysis also is valuable in detecting the causes of specific learning disabilities, as in reading. For instance, the test items may reveal deficient perception and recall of visual patterns, defective visual recognition, poor copying and reproduction of form, and short memory span—all of which are causes of confusion and difficulty in learning to read and to acquire other school subjects. (See Figures 15.1, 15.2, and 15.3). When such defects are revealed by the Stanford-Binet scale, a further intensive examination with a special reading test is indicated.

As illustrations of evaluations of Stanford-Binet performance, we cite the following reports:

Range of Allen's performance extends from the eight-year level to his CA level (10 years). His comprehension and reasoning with abstractions are at the level of his chronological age. Relative to his average level, S is generally superior in dealing with verbal abstractions. Comprehension of problems, analysis of absurdities, and problems of logical fact rank relatively high in his performance. He is weak in dealing with number concepts; cannot make change accurately. Visual perception and visual memory are inferior to chronological age expectancy. Memory for verbal materials is somewhat inferior to CA level. Vocabulary is inferior; S failed to pass at ten-year level; failed rhymes; word association is poor; reads with difficulty. (IQ about 90.)

² Compare, for example, with the Bellevue scales and the Chicago Tests of Primary Mental Abilities.

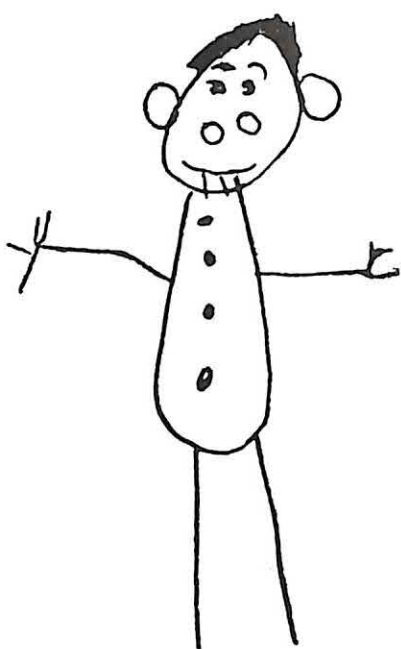
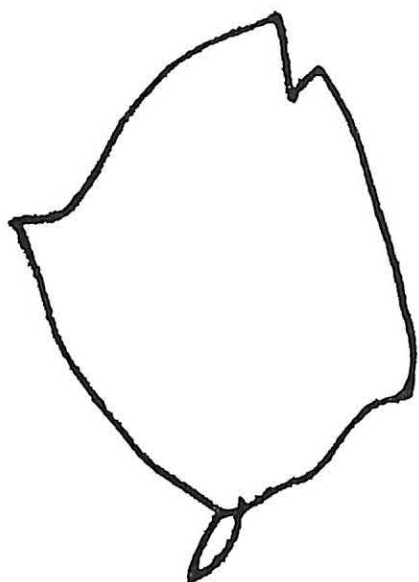


FIG. 15.1 (above, left). Copy of a diamond drawn by a boy with impaired visual-motor functioning. The same boy drew Fig. 15.2; CA, 9-5; Stanford-Binet MA, 8-0; IQ, 84. Compare this reproduction with Fig. 15.3.

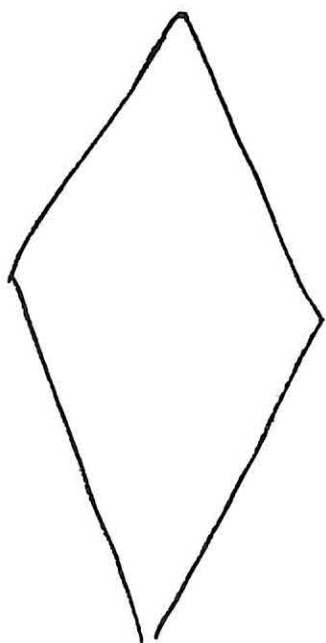


FIG. 15.2 (above, right). Drawing of the human figure by a boy with impaired visual-motor functioning. CA, 9-5; Stanford-Binet MA, 8-0; IQ, 84. This drawing, according to the norms of the Goodenough scale (Measurement of Intelligence by Drawings), gives this boy a mental age rating of 6-3, and an IQ of 66. See also Fig. 15.1.

FIG. 15.3 (left). Copy of a diamond by a mentally deficient boy. CA, 13-2; Stanford-Binet MA, 8-4; IQ, 63. Compare with Fig. 15.1.

Lois' test performance was not consistent. She did well on visual items requiring form discrimination and identification, but completely failed aesthetic comparison. She was unable to reproduce a square correctly, though hers was a close approximation; and she completely failed to grasp the concept of man-completion. However, she was able to copy folding of paper into a triangle and she did pass maze tracing. A consistent failure was that of memory—she could not remember either digit spans or sentences, though her speech difficulty may have been a major influence in the latter, especially. She seemed unable to grasp concepts as a whole; she often misunderstood directions and gave the impression that she occasionally made correct responses accidentally rather than as a result of understanding. She was self-critical on the drawing items, which she did painstakingly and with self-appraisal and subsequent erasures. She also was critical of her paper triangle and took another piece of paper after saying, "that's wrong" of her first attempt. She did not seem to realize her failures in other respects. (IQ about 55.)

Nancy was a very cooperative subject. She was a little apprehensive when she entered the testing room, but quickly relaxed and entered actively into the tasks. She had a long attention span and was not distracted by outside noises. She thought carefully before answering difficult questions, and was self-critical, for she sometimes thought aloud and corrected herself as she thought. She was rather self-confident and failure of items, some of which she recognized herself, did not seem to bother her. She laughed occasionally but on the whole was serious most of the time, giving all her attention to the test. (IQ about 125.)

Terman³ offers a psychological rationale which is helpful in interpreting results obtained with the Stanford-Binet scale, such as reported above. He gives, for example, the following: *ball and field*, practical judgment; *rhymes*, verbal associations under the direction of a guiding idea which inhibits irrelevant associations; *absurdities*, judgment and critical ability; *memory for design*, analysis of visual perception and creation of a whole which facilitates recall; *arithmetical reasoning*, ready and accurate application of knowledge to a given problem.

Scatter Analysis. A large number of investigations have been conducted in an effort to determine the extent of deterioration, or deficit, in intelligence suffered by each of the several groups diagnosed as

³ *The Measurement of Intelligence*, Boston: Houghton Mifflin, 1916.

having a certain form of mental disorder: for example, alcoholism, schizophrenia, manic depressive psychosis, paranoia, psychoneuroses, and others. One method commonly used has been to compare the averages of mental ages and the measures of variation found for a sample of abnormal persons with corresponding indexes found for a sample of normal persons. The theory is that inferior status of an abnormal group, as compared with one that is normal, is a result of the mental disorder. On the whole, it has been found that groups suffering from "organic" psychoses (e.g., alcoholism, paresis) show the greatest loss, or deterioration; those suffering from "functional" psychoses (e.g., schizophrenia, paranoia) show less; while those in the psychoneurotic group give no evidence of deficit. The published studies, however, do not agree as to the average amount of deterioration in each group. Furthermore, within each category of mental illness there are very marked individual differences in the degree of deterioration (as shown by measures of deviation).⁴

Due to these variations and to the overlapping of group distributions of ratings, the investigations of group deficits are of relatively little clinical value for individual diagnostic purposes. The investigations of group deficits have been valuable insofar as they have shown that some degree of deterioration of mental functioning accompanies most forms of mental disorder. Since clinical practice, however, is a matter of studying individuals rather than groups, it is necessary to devise means of determining individual rather than group losses of mental ability, or peculiarities of performance of individuals classified in each of the several categories.

One of these methods is the analysis of "scatter," which means the distribution of successes and failures over the several age levels, as in the Stanford-Binet scale. There are several methods of measuring scatter: (1) the number of age levels from and including the basal year up to the highest level where any items are passed; this is known as *range of scatter*; (2) the number of items passed above the individual's mental-age level and number failed below; this is known as *area of scatter*; (3) a combination of range and area whereby the number of successes and failures at each level are weighted according to the distance of that level from the mental-age level. If only one

⁴ For a concise summary and bibliography on this subject, see "Psychological Deficit," Chapter 32, by J. McV. Hunt and C. N. Cofer, in *Personality and the Behavior Disorders*, Vol. 2, edited by Hunt, New York: Ronald Press, 1944.

method is to be used, the first of these is preferable, since it covers the entire range of performance and expresses the situation in the most direct and simple manner.

Many investigators have reported that psychotic persons show greater scatter than do normal individuals or the nonpsychotic mentally deficient. Many reports state also that excessive scatter is a useful diagnostic sign, particularly in the case of "organic" psychoses which apparently yield the greatest degree of scatter.

In addition to general scatter, analyses have been made of "selective scatter"; that is, the effects of specific psychotic conditions upon ability to pass specific kinds of test items. For instance, schizophrenics are said to be less proficient than normal persons in detecting absurdities, interpreting fables, passing the purse and field test, memory for designs, and passing problem questions. On the other hand, they are said to be more proficient in arithmetic reasoning.

These views on both general and selective scatter are tentative. There are several reasons for disagreement among the published findings, one or more of which might operate in any given investigation: inadequate control groups, inadequate control of mental age and chronological age of groups being compared, divergent techniques of scatter analysis, and errors of psychiatric diagnosis.⁵

In view of inconsistent data, we must conclude that at present, numerical measures of general scatter on the Stanford-Binet scales are of limited use as clinical aids, so far as most individual cases are concerned. On the other hand, *extreme* degrees of scatter have been found diagnostically valuable often enough so that analysis of scatter on the Stanford-Binet continues as a clinical technique. Extreme scatter should make the examiner suspect one or more of the following: unusual extrinsic factors, emotional or organic disturbance.

Regarding selective scatter, there is a greater degree of agreement; for the published researches deal mainly with schizophrenics, among whom there are a number of different syndromes that affect test performance. Other abnormal groups have also been studied. Although investigators have not been in complete agreement regarding the several abnormal groups, all reports do show that schizophrenia does involve *selective impairment* of the mental functions tested. It

⁵ For a comprehensive bibliography on scatter, see D. Rapaport, *Diagnostic Psychological Testing*, Chicago: The Yearbook Publishers, 1945, pp. 554 ff.

is to be expected that the particular functions impaired, and the degree, will vary with the particular type of schizophrenic disorder; but in this group as a whole test performance tends to be inferior more often and more markedly in parts requiring practical reasoning and judgment (What is the thing to do when . . . ?), in abstract reasoning (arithmetic problems), and in perceptual organization. Inferior performance is revealed not only in the lower scores but in the quality (e.g., bizarreness, irrelevance) of responses and inconsistencies (failing easy items and passing more difficult ones).

The Stanford-Binet has been used, among others, in making inter-test comparisons of verbal with performance scales, to discover if the scatter or *psychometric pattern* of each kind of test might have diagnostic and differentiating value with atypical groups and individuals.⁶ These comparisons have shown that psychotic persons and habitual criminals are less proficient on performance than on verbal tests; their best scores are obtained on tests of vocabulary, next highest on the Stanford-Binet, and lowest on performance scales.

In the cases of most delinquents and mental defectives—non-psychotic atypical groups—a quite different order or pattern of performance was found. Their vocabulary and Stanford-Binet scores were low, while their performance test ratings were higher. Some individuals in these two groups, however, had relatively low performance test ratings; but they, it was found, were least likely to make satisfactory social adjustment.

It has been reported, also, that among maladjusted children most of those with relatively very high performance test ability were individuals referred to clinics as "delinquents"; while children with significantly lower performance than verbal ratings were mostly those having personality defects (psychoses, neuroses, emotional instability).

Published studies are in general agreement that individuals with relatively and significantly inferior performance-test ratings present most of the behavior problems, are most unstable, and show least promise of adjusting satisfactorily in a training school.

⁶ See S. W. Bijou, "The Psychometric Pattern Approach as an Aid to Clinical Analysis—a Review," *American Journal of Mental Deficiency*, Vol. 46, 1942, pp. 354–362; S. W. Bijou and B. R. McCandless, "An Approach to a More Comprehensive Analysis of Mentally Retarded Pre-Delinquent Boys," *Journal of Genetic Psychology*, Vol. 65, 1944, pp. 147–160.

Why these psychometric patterns should prevail has not yet been explained satisfactorily. Poor "behavior efficiency" has been suggested as one explanation of the pattern found with psychotics and habitual criminals. This notion, however, offers little help unless it is interpreted to mean a loss in capacity to make adaptations to new or unfamiliar situations—such as would be encountered in performance tests. The test pattern of these two atypical groups might be due to lack of sustained attention and reduction in speed of response, both of which generally are demanded more on performance than on verbal scales. Verbal tests, on the other hand, particularly at the adolescent and adult levels, draw heavily upon information, knowledge, vocabulary, and habits of thinking, all of which have become rather well established, if not entirely fixed, and which, therefore, are likely to become less impaired during mental illness or periods of emotional instability.

The fact that the psychometric pattern for the mentally deficient group shows their performance-test ratings to be higher than their verbal abilities should not occasion surprise, for studies of this group have repeatedly shown that educationally and psychologically one of their major deficiencies is in their capacity to deal with language and number concepts (abstractions in general as compared with the concrete). Why delinquents should reveal a similar pattern is uncertain. An explanation may be found in some or all of the following: delinquents are, as a group, at a lower intellectual level than an unselected population; they are, as a group, more likely to come from disadvantaged homes, poorer cultural and educational backgrounds; they may be lacking in motivation on verbal tests.

There are, finally, some important statistical facts to consider in connection with interpretation of scatter on the Stanford-Binet. The first is that the percents of the standardization population passing each item at any given age level are not all equal. For example, at age 7 (Form L), the percents passing vary from 51 to 70; at age 8, from 57 to 67; at age 12, only from 61 to 69. These differences signify that the items at a given age level are not all of equal difficulty; hence uniformity of performance is not to be expected. It will be recalled that placement of items was made on the basis of several considerations, only one of which was percent passing.

Other technical factors that might affect scatter are the reliability of each item; or low intercorrelations between certain items, some of

which might make demands upon a more or less specialized ability (low correlation with g); or items wrongly placed in the age-scale. Though these technical factors may account for part of the scatter found, they cannot account for all of it, nor for whatever degree of agreement exists in the findings for a single clinical group.

Although the psychological hypotheses offered to account for some scatter-patterns are only tentatively advanced, knowledge of the existence of these psychometric patterns in some cases is clinically valuable and emphasizes an approach that often has been useful and that, therefore, continues to be used.⁷

Content Analysis. In recent years, clinical psychologists have been increasingly interested in analyzing verbal responses to test items for their possible significance regarding the examinee's experiences and personality traits. In other words, the Stanford-Binet, and other individual scales, may have some value as projective tests (discussed in a later chapter).

In the Stanford-Binet, the number-of-words-in-one-minute item (Form L, year X, item 5) may reveal current interests, family conditions, school relationships, etc. One ten-year-old boy (IQ 133) interpolated the following words in his list: *school, ugh, arithmetic, dull, stupid*. In spite of his high mental level, he was having difficulty with his studies (particularly arithmetic) and with his teacher. That the school was a source of anxiety to him was clearly revealed in a sentence completion test and in the Thematic Apperception Test. Other psychologists report words of violence, ambition, unsatisfied needs, etc. Verbal responses may thus reveal more or less deeply rooted emotional conditions.

Word definitions may reveal experiences and centers of interest. For example: as when "puddle" is defined as "dirty water that you should not go into"; or "lecture" as "a lot of yak-yak from a grown-up." There are also times when definitions will be concentrated in a

⁷ On scatter, see also: M. M. Riggs and K. A. Burchard, "Intra-Scale Scatter for Two Kinds of Mentally Defective Children," *The Training School Bulletin*, Vineland, N. J., Vol. 49, 1952, pp. 36-44; A. Magaret and C. W. Thompson, "Differential Test Responses of Normal, Superior, and Mentally Defective Subjects," *Journal of Abnormal and Social Psychology*, Vol. 45, 1950, pp. 163-167; W. Sloan and R. A. Cutts, "Test Patterns of Mental Defectives on the Revised Stanford-Binet Scale," *American Journal of Mental Deficiency*, Vol. 51, 1947, pp. 394-396.

special area (e.g., sensory) so as to suggest special interests. Some give definitions attended by compulsiveness, uncertainty, or over-elaboration.

Values and behavior patterns also are revealed at times. For example, under "Comprehension" (Form L, year 7, item 4), one of the questions is this: "What is the thing for you to do if another boy hits you without meaning to do it?" A common, and expected, answer is: "Nothing. He didn't mean it." But at times a boy replies: "Hit him back" or "Go home and tell my mother (or teacher)."

Picture interpretations also may offer an opportunity to gain insights into aspects of behavior and personality. The responses to pictures may be moralistic, or submissive, or hostile, or anxious, etc. Or they may be of the ordinary, expected variety.

Some children who give ready and assured responses to routine materials, like recall of digits and sentences, may be anxious, hesitant, and tentative when dealing with items that require their own judgment and evaluation, thus suggesting dependence or submissiveness.

In the test situation it is possible for experienced examiners to utilize certain practices that have not been prescribed in the manual, nor even contemplated by the test's author. As instances in point, the three following practices are cited. After the formal testing has been completed, it is sometimes desirable to check on certain failures to determine whether they were due to lack of capacity, verbal handicap, or failure to understand instructions.

This may be done by giving *actual instruction* on certain types of items, then testing the individual further on the same types. The results, of course, do not change the subject's score on the test; but the procedure and its results should be noted in the report. For example, if the child has failed on similarities and differences (Form L, year VIII, item 4), when testing has been concluded the examiner might return to that item, explain the terms and the nature of the problem, and even give the correct answer to one of the four parts. He then asks the subject to give the answers to the remaining parts to determine if there has been any learning.

A second kind of "extra-testing" procedure is directed toward discovering how far a testee has been able to proceed on an item he has failed, as in arithmetical problems. Thus, the examiner returns to the problem; he leads the testee through the solution step by step, at times suggesting the procedure or even the basic process to be

used. In this manner, it is often possible to discern the nature and extent of a subject's deficiencies.

If a child has failed to copy the diamond satisfactorily (year VII, item 3), the examiner may want further information on whether the inability is probably due to disturbed visual-motor functioning or to defective perceptual analysis. He may, then, make three copies of the diamond, of varying quality, one of which is a duplicate of the testee's own copy. The testee is then asked to select the "best" one. It has often been found that the child whose poor copy is consistent with his general mental level will not select the examiner's superior reproduction or will select at random; but a subject whose general mental level is higher than that suggested by his own copy can make the correct selection, even though he still will not be able to produce a satisfactory copy on subsequent trial.

These are three illustrations of "extra-testing" procedures that enable the examiner to prepare a more analytical and richer report.

Order of Item Administration. The order in which items of the Stanford-Binet should be presented has been under examination and discussion in recent years because clinicians have found that order of presentation in some instances influences performance of maladjusted subjects. Three methods have been used: (1) standard, or consecutive; (2) serial; (3) adaptive. Using the first of these, the examiner presents the items in the order in which they appear in the scale. Using the second, he follows through on *one type* of item, to the more difficult ones, until the subject fails: for example, memory span for digits, memory for sentences, similarities and differences, comprehension, etc. If the third method is used, the examiner starts with easy items and alternates these with difficult ones of the same type. When employing the adaptive method, the examiner begins at a level below the subject's expected mental age (to insure success) and moves up and down in an effort to establish the maximal (or terminal) and the basal levels as early as possible. The principal justification given for the adaptive method is this: some individuals are discouraged by a series of consecutive failures when they reach higher and more difficult levels; they might, therefore, fail items that they could otherwise pass. The adaptive method is intended to scatter failures among successes.

The sparse experimental data support the use of the adaptive pro-

cedure. First, it has been found that there is little difference between results found by standard and adaptive methods when a representative population or a well-adjusted group are tested. Thus the norms established by Terman and Merrill through standard testing are applicable when the adaptive method is used. One study reports a correlation of .93 between results obtained when the two methods were compared, using Forms L and M. The test-retest differences between IQ's were just about what Terman and Merrill reported in their standardization data: namely, about 5 points.

Second, with the adaptive method, *maladjusted subjects*, especially the more maladjusted, achieve higher IQ ratings than they do by the standard order of testing. One study reports a mean gain of 11 points when the adaptive method was used, as compared with the standard consecutive method.⁸

Insofar as an adapted method of procedure contributes to a fuller and more valid understanding of a subject's performance, without violating the basic principles of the scale being used, its use in appropriate cases is warranted.

THE BELLEVUE SCALE

For purpose of analysis, this scale has an obvious advantage over the Stanford-Binet because its items are grouped into eleven subtests, each examining psychological functions which might be susceptible to different forms of personality maladjustment. Arrangement into subtests not only permits the preparation of analytical profiles for individuals, but it makes possible the ready analysis of subscore interrelationships which may be studied for any bearing they might have on clinical use.

Clinical Diagnosis. Since 1940 a rather large number of studies have been reported, detailing differences found between normal and abnormal groups, and between the several abnormal groups themselves. These researches were concerned with differences in extent of scatter and with differences between patterns of subtests, emphasis being placed upon the latter. There seems to be agreement, for example, that

⁸ A. N. Frandsen *et al.*, "Serial *versus* Consecutive Order Administration of the Stanford-Binet Intelligence Scales," *Journal of Consulting Psychology*, Vol. 14, 1950, pp. 316-320; M. L. Hutt, "A Clinical Study of Consecutive and Adaptive Testing with the Revised Stanford-Binet," *ibid.*, Vol. 11, 1947, pp. 93-103.

schizophrenics are relatively superior on the subtests of Information and Comprehension, while they are inferior on Object Assembly and lowest on Digit-Symbol. One investigator, Rabin, devised a "Schizophrenic Index": namely, the ratio of the sum of the scores on Information, Comprehension, and Block Design to the summated scores on Digit-Symbol, Object Assembly, and Similarities.⁹ The justification for this empirical index was that it seemed to differentiate between schizophrenics, on the one hand, and manics, and normals, on the other. This index, however, must be regarded as tentative.

The foregoing studies were devoted principally to the determination of group trends and group patterns. Results obtained from study to study were not always in agreement; nor were the results of a particular report always unequivocal. We mention these investigations because they are efforts to make the Bellevue scale more valuable analytically and clinically; and, as tests are improved, it is possible that these methods may yet prove valuable as analytical techniques.

Brown, Rapaport *et al.*, first suggested the *intraindividual* measure of scatter.¹⁰ This consisted of: (1) calculating, for each individual, the deviation of *each* subtest score from the *mean of all* his subtest scores; (2) computing the deviation, for each individual, of each verbal subtest score from his verbal subtest mean; (3) computing the deviation of each performance subtest score from the performance subtest mean. This approach proved to be promising and served as the preliminary study to an extensive and detailed investigation to be reported below.

The author of the Bellevue scale took note of the analytical and clinical investigations reported by other psychologists, added his own clinical findings, and prepared a table giving the test characteristics, or patterns, of several clinical groups: organic brain disease, schizophrenics, neurotics, adolescent psychopaths, and mental defectiveness.¹¹ As an example, according to this table, the pattern of the neurotic group shows *relatively* superior scores on Information, Comprehension, Similarities, and Vocabulary; relatively low scores on Digit Span, Picture Arrangement, Object Assembly, and Digit-Sym-

⁹ A. I. Rabin, "Test-Score Patterns in Schizophrenia and Non-Psychotic States," *Journal of Psychology*, Vol. 12, 1941, pp. 91-100.

¹⁰ See D. Rapaport, *op. cit.*, p. 553.

¹¹ See D. Wechsler, *The Measurement of Adult Intelligence*, 1944, Chapter 11.

bol; average scores on Arithmetic, Picture Completion, and Block Design.¹² The explanation of this and other patterns is, presumably, that mental defect, each major type of personality disorder, and mental illness are associated with characteristic deficits or losses of psychological functioning. Mental defectives are individuals whose mental development has been *arrested* at a relatively low level, though not uniformly in respect to all functions. A personality disorder or a mental illness produces impairment in the individual's mental abilities; but the resultant losses are not uniform; apparently some functions are more adversely affected than others. Thus, if a psychological test is able to measure and portray these differential deficits and losses, it can assist significantly in making a clinical diagnosis.

We must, however, emphasize a caution in this connection. The test pattern of each group is not entirely unique; that is, the same types of tests are generally failed by more than one clinical group, and still other types are generally high for more than one group. For instance, psychopaths and mental defectives score relatively better on nonverbal than on verbal subtests; in general those suffering from organic brain disturbances, schizophrenia, or neuroses score relatively low on Digit-Symbol and Arithmetic. There are also some differences between these clinical groups, as well as similarities. The tests cannot be used as a short-cut to clinical diagnosis; but for one who is familiar with the general character of the psychological functions being tested and which have been impaired, analysis of an individual's performance on the Bellevue scale, taken together with other evidence, adds to the probability of a correct diagnosis of the case.

Scatter Analysis. The most detailed report on the Bellevue scale as a diagnostic instrument was presented by Rapaport.¹³ His study is based upon the examination of 217 clinical cases and 54 control cases. The control group consisted of "54 randomly chosen members of the Kansas Highway Patrol." This control group is the weakness of Rapaport's research; for it is doubtful if a small number of Kansas Highway Patrolmen constitute an adequate control against which test patterns of patients should be compared. Furthermore, in order to evaluate the

¹² The reader should note that in a particular instance each subtest score is high, low, or average with reference to an *individual's own* average score, not with reference to average performance of any group.

¹³ *Diagnostic Psychological Testing*, Vol. 1.

test patterns of the members of the control group it was necessary to establish an adjustment rating for each one. If an individual was well-contained, a satisfactory worker, without evidence of maladjustment beyond the ordinary occasional difficulties, anxieties, or mood fluctuations, he received a rating of 1. Those whose behavior was considered borderline or maladjusted—in respect to instability, childhood symptoms, impulsiveness, moodiness, etc.—got a rating of 2. Patrolmen who were clearly maladjusted were rated 3. These ratings were based upon two hours of interview and social, developmental, and occupational records.

We may question not only whether the group itself is adequate for control purposes, but also whether the evaluations and ratings of the patrolmen's personalities rested upon a firm enough basis. Nevertheless, Rapaport's report on the Bellevue scale is the most exhaustive and resourceful one yet to appear. It indicates the clinical possibilities of an intelligence test; it attempts to provide the psychological rationale of each subtest, on the basis of which the nature of mental deterioration or deficit may be more analytically and precisely represented. We shall, therefore, briefly present Rapaport's method, with several illustrations for clarification. The method employed was scatter analysis, essentially in two forms.

(1) *Vocabulary scatter.* This is the difference between a person's score on a particular subtest and his score on the Vocabulary subtest. The reason for using Vocabulary as the base of comparison is that rather consistently it has been found to be the psychological test least vulnerable to impairment by personality maladjustment or mental disturbance. It is the Vocabulary score, therefore, from which the individual's original, unimpaired intelligence level can best be inferred. Degree of loss in other functions can thus be derived from the differences between ratings on each of the subtests and the rating on Vocabulary.

(2) *Mean scatter.* This is the difference between the rating on any single subtest and the average of the ratings on all the remaining subtests.¹⁴ The mean scatter, which may be positive or negative for a

¹⁴ In calculating mean scatter, Rapaport omitted Digit Span and Arithmetic. This omission, he states, ". . . was warranted by the fact that impairments [of these functions] were so general in most of the clinical and control groups, that their inclusion would have vitiated the representativeness of the mean as a central tendency of the scores." (*Op. cit.*, p. 53 n.)

given subtest, measures the relationship of a single measured function to the average of all other functions measured by the test. It is thereby possible to find out whether an individual's level of achievement and functioning on any of the subtests has deteriorated more or less than on the remaining ones.

Mean scatter can be calculated for the entire test, including both verbal and nonverbal (performance) subtests, or it can be found separately for the six verbal scores and the five nonverbal scores.¹⁵

A third procedure—one which supplements scatter analyses and is regarded as essential—is the comparison of *pairs* of subtest scores. Any two subtest scores may be compared for the purpose of finding out how the subject's functioning in one compares with his functioning in another. For example, how does his retention of information (remote learning) compare with his memory span for digits (immediate recall and attention)? How does his arithmetic score (reasoning and habitual responses) compare with his score on similarities (concept formation)?

Thus, it is possible to obtain in considerable detail an analysis of the subject's achievement levels on the several parts of the scale, their interrelationships, and evaluations of general and special mental impairments. Scatter on the Bellevue scale, it is maintained, “. . . is not random, but follows definite rules and is diagnostically differential between kinds of clinical and normal groups.”¹⁶

The profile in Figure 15.4 illustrates the method we have been discussing.¹⁷ We quote, also, from the interpretation of this case not only to clarify the profile but to emphasize the fact that mechanical scatter analysis is inadequate and to show how the psychological rationale of the scale may be utilized in interpretation.

It is true, unfortunately, that not all clinical cases present clear-cut patterns of test performance upon which a differential diagnosis may be made. In an appreciable number of cases, scatter and subtest analyses have been found definitely diagnostic; in other cases, the analyses, though not conclusive, provide indications of the probable

¹⁵ Rapaport describes three methods of scatter analysis: verbal, mean, and modified mean. The last two are so much alike that they have been combined to show the principle involved.

¹⁶ Rapaport, *op. cit.*, p. 54.

¹⁷ From R. Schafer, “The Expression of Personality and Maladjustment in Intelligence Test Results,” *Annals of the New York Academy of Sciences*, Vol. 46, 1946, pp. 609–623. This report is based upon studies directed by Rapaport.

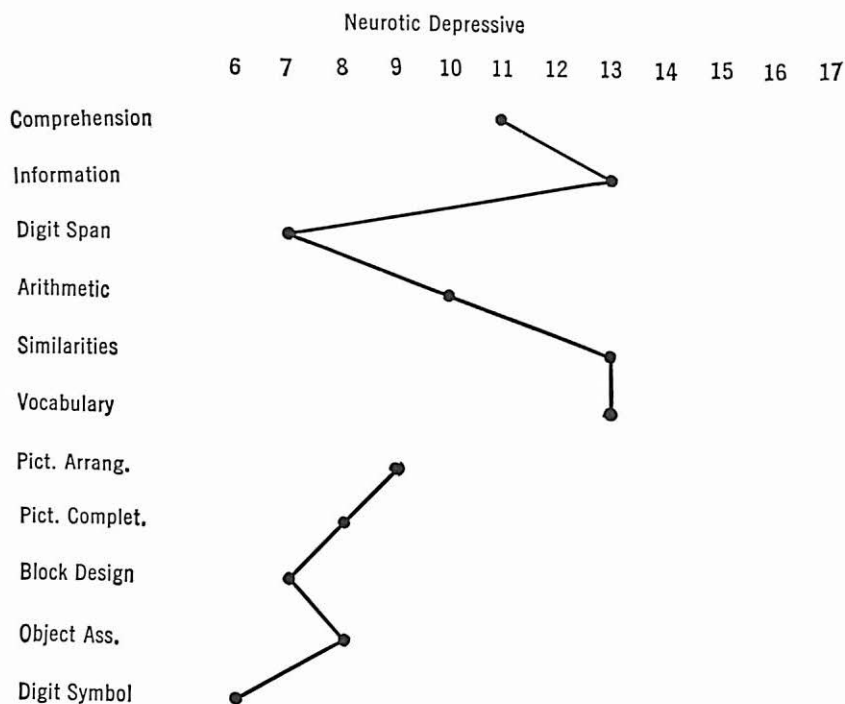


FIG. 15.4. "Outstanding in the scatter is the great discrepancy between the verbal and the performance sub-test scores, and this discrepancy we have found to be a statistically significant, and therefore diagnostic, indication of depression. The rationale of this finding is the following: depression becomes manifest in intellectual functioning by a retardation of perceptual and associative processes. The relatively complex visual organization and visual-motor coordination required by the performance subtests put too great demands upon the slowed-down depressive. Furthermore, in contrast to the untimed verbal subtests, the performance subtests have time limits on each item and even give extra credit for speed. Consequently, depressives not only do not obtain extra credit, but exceed the time limit on many items. For this case, item analysis confirms the retardation by showing that, on Picture Completion and Block Design, a number of items were failed only because they exceeded the time limit. . . . [t]he impairment of Digit Span or attention is also striking and reflects the presence of intense anxiety accompanying the depression. The mild impairment of Arithmetic is referable to an inability to meet the time-limits and gain time-bonuses on the items of this subtest. In verbalization, much self-depreciation, as well as indirect criticism of the test and the examiner, are evident." R. Schafer, *op. cit.* (By permission.)

diagnosis; in still others, the scatter profile and analyses are quite inconclusive. In spite of the fact that Bellevue test-analysis does not provide a definitive diagnosis in all cases, such analysis is warranted because it is practically as efficient as other diagnostic procedures. Intelligence test analysis should never be used alone or independently; it should be used and interpreted in conjunction with the results obtained with other psychological tests (e.g., projective methods) and other clinical and developmental evidence.

One specific precaution must be emphasized in connection with inferences to be drawn from scatter analysis of the Bellevue, the Stanford-Binet, or any other intelligence test; namely, a subject may manifest certain deficiencies on a test due to educational, cultural, or other environmental deprivation. On the other hand, environmental predilections or pressures may enhance certain forms of test achievement. In either instance, the diagnostically distinguishing features of the intelligences scales will be affected and probably invalidated. Thus, before drawing diagnostic inferences from intelligence test results, the psychological examiner must know whether, in the case under consideration, there have been educational, cultural, or other environmental factors that might account for any of the ratings and seemingly diagnostic criteria obtained.

Qualitative Aspects of Responses. Scatter analysis and subtest comparisons should be accompanied by a qualitative analysis of the subject's verbalizations and general approach to the test problems. Such an analysis reveals characteristics of the individual's mentality in operation: excessive doubt, indecision, self-criticism, impulsiveness, bizarre notions, obsessions, random guessing, etc. Certain of these and other qualitative characteristics have been found frequently associated with clinical groups.

In view of the lack of uniformity of subtest patterns for the several clinical classifications, a testee's mode of approach to a test item is highly significant, psychologically, in the interpretation of behavior and in the understanding of his personality. Classification and labeling are less important, in an individual case, than description and analysis of that person's behavior and functioning, as shown by subtest results.

A person's mode of approach will affect his performance differently on each of the various types of subtests, depending on whether the

subtest makes demands, for example, on habituated responses or on responses requiring flexibility and reorganization. Individuals with brain damage—as a case in point—when tested show rigidity, inability to shift attention, inability to change their mode of responding, inability to ignore superficial or extraneous stimuli, and difficulty in organizing material into a pattern or a meaningful logical sequence.¹⁸ Inspection of the Bellevue subtests will show that performance on them would be variously affected by behavioral traits such as these.

Failure to respond correctly to a test item is not the only matter of importance to the clinical psychologist. While correct and acceptable responses to test items are fixed by research and by agreement of psychologists regarding the interpretation of the findings, the incorrect responses are not fixed; nor is the *manner* of responding fixed. To illustrate: current tests have sets of items testing “comprehension,” “similarities,” and “arithmetical reasoning.” The first are of the “when” or the “why” type; that is, “what is the thing to do when . . . ?” or “why is it desirable (or necessary) to . . . ?” The second, “similarities,” requires understanding of basic likenesses—for example, “In what way are _____ and _____ alike?” The arithmetical problems range from the simplest to fairly complex. A person’s response to any of these may be right or wrong; but his speed and confidence in answering, or his anxiety or blandness about incorrect responses often reveals significant personality traits. When, in a test of information, an otherwise intelligent person blandly replies that Tokyo is in Turkey, or, in a test of similarities, that a dog and a lion are alike because both have digestive organs, and gives other bizarre responses, personality disturbance is strongly indicated.

Personality disturbance of the obsessive kind, for example, must be considered when a subject feels compelled to offer four or five explanations of courses of action in reply to a “comprehension” item; or when he mentions three, four, or more likenesses on some of the “similarities” items; or when he gives elaborate and often quibbling definitions of words in the vocabulary test. Or if the person persists in guessing blindly on test-items that are clearly beyond his level of ability, his test-behavior may be indicative of an uncritical impulsive-

¹⁸ Cf. R. Goldman et al., “Use of the Bellevue-Wechsler Scale in Clinical Psychiatry with Particular Reference to Cases with Brain Damage,” *Journal of Nervous and Mental Diseases*, Vol. 104, 1946, pp. 144-179.

The foregoing are illustrations of the role of language with respect to certain *types of items*. Some personality patterns or categories may, however, be inferred from the manner in which the person deals with the test in general and as a whole. A few illustrations follow.

In the "obsessive-compulsive" individual, verbalization of responses is over-detailed and doubt-laden. For example, in response to the question: "What does 'stanza' mean?" the ordinary person who has the information would probably say, "A group of rhymed lines," or something similar. A characteristic obsessive-compulsive response, however, would be comparable to this: "A stanza of rhymed lines forming one of a series of similar divisions in a poem. Two rhymed lines form a couplet. A four-lined stanza is a quatrain, a six-lined one is a sestet," etc. One or a few such replies do not warrant the characterization of the respondent as obsessive-compulsive; but when this kind of answer is "idiomatic" of the individual, such characterization is indicated.

Persons in an "anxiety state" also frequently give responses that are typical of that group. In general, their behavior is characterized by restlessness, apprehensiveness, impaired attention and concentration, and bodily expressions (such as tics, nail-biting, fidgeting, coughing, etc.). This psychological state is manifested, in the test-situation where language is required, through difficulty with finding words, impulsively blurting out unfinished or unchecked or inappropriate replies, or fumbling about for adequate formulations. For example, when the question is: "How many weeks are there in a year?" the anxiety-ridden person may reply: "There are 48 weeks in a year; no . . . let's see . . . or is it 50? . . . wait a minute . . . let's see . . . 12 months . . . 4 weeks in a month . . . yes, that's right, 48." Or consider the question: "Why should we keep away from bad company?" The person in an anxiety state may give a reply of the following sort. "If I was in bad company, I'd get away quick. I wouldn't want to be with them in the first place. They . . . er . . . are unlawful . . . and . . . ah . . . get you in trouble. . . . I don't think a person should be in bad company; that is, . . . er, ah . . . if he was brought up right. Anyhow, I'd leave them!"

The test responses of psychotic persons, also, are quite significant for diagnostic purposes; but the subject of psychotic responses is a complex one. One aspect of the responses of psychotics may be indicated here: namely, disorganization of thinking and bizarreness of

responses, typical of schizophrenics—both, of course, being indicated in the language of their answers. For example, the former teacher of history who cannot correctly give Washington's birthday, or the clergyman who cannot define the word "vesper" are instances that indicate disorganization of memory and loss of previous knowledge. Or, on the "bad company" question, when one gives an emotionally intense and moralistic response, and explains, irrelevantly, why people should be "good." Such a response suggests serious impairment of judgment. Or in the vocabulary test, when a subject gives impulsive replies, such as defining "belfry" as "a kind of bellboy"; or "repose" as "to pose over something." These bizarre answers indicate an impulsive "clang-association."

Deterioration Index. The Bellevue scale presents a procedure for calculating the approximate amount of mental deterioration due to advancing age.¹⁹ This index rests upon the principle that some of the mental functions being tested decline more rapidly than do others. The difference between rates of impairment, as shown between the two types of functions, is said to indicate, in the case of any particular individual, his degree of deterioration. There are certain tested functions that "hold up" with age whereas others do not. The subtests reported as "holding up" with age are: Information, Comprehension, Object Assembly, Picture Completion, Vocabulary. Those that do not "hold up" with age are: Digit Span, Arithmetic, Digit-Symbol, Block Design, Similarities, Picture Arrangement. Since some loss of tested abilities is normally expected with advancing age, this factor is taken into account in calculating the deterioration index.

As a first step it is necessary to calculate the average loss due to deterioration: that is, the total of "hold" scores minus total of "don't hold" scores, divided by the total of "hold" scores. The empirical formula is: ²⁰

$$\text{Deterioration Index} = \frac{\text{Hold} - \text{Don't Hold}}{\text{Hold}}$$

For example, assume that a man 35 years of age has the following scores: "hold," 50; "don't hold," 35. Substituting these values in the

¹⁹ Wechsler, *op. cit.*, Chapter 6.

²⁰ This empirical formula, it is clear, uses the sum of the "hold" scores as the basis of comparison; the larger the difference between it and the "don't hold" scores, the greater the apparent deterioration.

formula, the deterioration index is 30 percent. Since only a 5 percent loss is expected at this age (according to Bellevue norms), this man's excess or net loss is 25 percent, which is regarded as significant and as abnormal deterioration.

The reader will recall that in the standardization of the Bellevue scale, the age norms for each subtest allow for declining scores in these measured functions after the age of their measured maximal level. By summing the means of the "hold" tests and the means of the "don't hold" tests at successive age levels, it is possible to estimate the average or normative loss on test scores to be expected with increasing age. The difference between the loss shown by an individual and the normative loss of his age group may then be calculated to determine whether that individual's deterioration index is greater or less than expected for his age. If a significant discrepancy is found, then the reasons and significance thereof should be sought.

The principle of the deterioration index is presented here as an aspect of the Bellevue scale, but not because its soundness has been unequivocally demonstrated. In fact, published researches on the index differ: some support the concept; others do not; still others are equivocal. One of the difficulties, again, seems to be that the conception of "hold" and "don't hold" subtests might apply to some age groups and some clinical groups, but not to others; whereas with most persons, selective rather than over-all loss might be the case.

The soundness of this method will depend, also, upon high reliability of the test, and upon the adequacy of standardization at all age levels. The standardization process for the Bellevue has been described in detail; and its weaknesses have been indicated. It appears, therefore, that the deterioration index is useful primarily in those instances where loss is so marked that imperfect test reliability and inadequacy of standardization cannot account for the discrepancy. Wechsler states that the method has proved effective in practice and may be applied with "reasonable confidence" if the test's limitations are given due consideration. Further experimental validation of the method is necessary.

For the present, comparisons with premorbid test results (when available) would be preferable as evidence of loss of mental efficiency. Where such data are not available, the index is of limited value and should be supplemented by inter-comparisons of subtest scores and by analysis for internal evidence of loss.

KENT SERIES OF EMERGENCY SCALES²¹

These scales, commonly referred to as the Kent E-G-Y, are intended for clinical use when the "examiner is . . . in need of a simple and informal mental test that can be presented very briefly as a preliminary measure." Performance upon this scale, it is held, is a useful criterion in deciding at what mental level to start the more formal and thorough examination.

The series consists of four overlapping scales, each with its own norms: Scale A, ages 5-7; Scale B, ages 6-8; Scale C, ages 7-10; Scale D, ages 9-14+. Each is very brief, requiring only ten to fifteen minutes. The scales consist only of a series of questions on objects and matters of common occurrence. The following are examples.

Scale A:

- Which is larger, a cat or a kitten?
- At what time of year is it very cold?
- What do we use our eyes for?

Scale B:

- What is a key used for?
- What does the cow give us?

Scale C:

- What does a baby chick come from?
- What makes it light on a cloudy day?

Scale D:

- What are houses made of?
- If your shadow points to the northeast, where is the sun?

Evaluation. The original Kent scale was devised for clinical use principally with delinquents and patients in mental hospitals, as a preliminary screening instrument. During World War II, it was modified and shortened for emergency screening. Since then, the newer scales have been constructed. Published data on the value of these emergency scales are equivocal. Some report high correlations with the

²¹ By Grace H. Kent. Published by The Psychological Corporation, 1946. See by Kent, *Oral Test for Emergency Use in Clinics*, Mental Measurement Monographs, Serial No. 9, 1932, Williams and Wilkins; F. A. Mullen, "Comparison of the Revised Kent Emergency Tests with the Revised Stanford-Binet and the Kuhlmann-Anderson Tests," *Journal of Psychology*, Vol. 15, 1943, pp. 151-163; H. P. Hogan, "Comparison of Stanford-Binet and Kent Oral Emergency Scale," *Journal of Genetic Psychology*, Vol. 58, 1941, pp. 151-159; W. A. Hunt and I. Stevenson, "Psychological Testing in Military Psychology. I. Intelligence Testing," *Psychological Review*, Vol. 53, 1946, pp. 25-35.

Stanford-Binet, but, at the same time, appreciable IQ differences. Others report only low or moderate correlations. Practically all seem to agree, however, that the scales rate superior subjects lower than does the Stanford-Binet, while mentally retarded subjects are rated higher. It has also been found that a large proportion of persons with behavior disorders are ranked significantly lower on the Kent than on the Stanford-Binet. This offers a suggestion for possible use of these two scales with some clinical cases.

Any emergency, quick-screening test necessarily must be relatively coarse; and as such its results will be much less reliable than those obtained with longer, more varied, more penetrating tests. The use of a very brief test is warranted only in an emergency. This is emphasized by Kent, who intended the scales to be employed only as one unit in a battery. They are not intended as a complete examination and should not be used as such.²²

A REPORT OUTLINE

The following is an outline used in teaching graduate students to prepare reports of tests of intelligence administered by them under supervision. This report outline will help to clarify the weight and significance given to qualitative aspects of test performance.

PSYCHOLOGICAL EXAMINATION

Name: (last, first, middle) Date:
Age: (years and mos., e.g., 8-7) Date of Birth:
Referred by:

Development of the report should follow this general outline:

- I. Introductory statement
 - a. By whom tested
 - b. Why tested
- II. Give a summary statement of the child's (or adult's) reaction to the test situation, the examiner, etc.
- III. State the name of the test given, *in full*.
- IV. Test results indicated by:

²² A useful summary of psychological examining in clinical situations will be found in J. W. Carter, Jr. and J. W. Bowles, Jr., *A Manual of Qualitative Aspects of Psychological Examining*, Clinical Psychology Monographs, No. 2, 1948.

*Intelligence Tests as Clinical Instruments**Stanford-Binet*

- a. Mental Age
- b. Intelligence Quotient
- c. Basal Age
- d. Terminal Age
- e. Descriptive Classification
(average, superior, etc.)

Wechsler-Bellevue

- a. Verbal IQ
- b. Performance IQ
- c. Full Scale IQ
- d. Descriptive Classification

V. Test Evaluation

- a. Indicate the performance of the person with regard to:
 - (1) verbal material
 - (2) non-verbal material
- b. Compare MA (or IQ in W-B) with achievement on Vocabulary test score
- c. Point out quantitative and qualitative aspects of test
 - (1) point out strengths
 - (2) point out weaknesses
- d. State your evaluation of the performance of the person

VI. Summary statement

The quantitative analysis can be derived from knowledge of which psychological functions the various test items are sampling, and by comparisons of how well the subject handles the various types of test items.

The qualitative evaluation may be derived from the following:

1. Reaction time: Were responses delayed, blocked, irregular?
Was there any indication of negativism?
Were the responses given quickly or impulsively?
2. Responses: Are they nonsensical, immature, childlike?
Are they good; are some better than others?
Is there confabulation?
Are there peculiarities of speech?
3. Depth of responses: Are they popular and surface responses?
Is the subject profound or does he attempt to be profound? (Judge this by comparison with information about subject as well as by content of responses.)
4. Self references: Is the question (or the answer) referred to the self?

Are the responses in terms of his own or immediate experiences or in terms of someone else's experience?

5. Evidence of confusion or doubt: Do questions have to be repeated?
Does the subject change his answers?
Is the question misunderstood or misinterpreted? If so, in what way?
6. Verbalization: Is the subject verbose?
Is he spontaneous in responding?
7. Organizational methods: Is he careful or over-meticulous?
Does he plan readily?
Does he generalize readily?
Is there evidence of perseverance?
8. Adaptability: Does he shift rapidly from one test to another?
What is his level of interest?
9. Motor coordination: Are fine and gross movements skillful or awkward?
Can he execute complex bilateral movements?
10. Effort and concentration: Is he cooperative?
Does he try hard?
Does he have trouble in attending?
11. Mood: Is he easily upset, irritable, emotional, argumentative, stuporous, elated, happy, sad or depressed?
Does his mood change during the testing situation? ²³

²³ I am indebted to my former graduate student and assistant, Dr. Joanna Byers, who collaborated on the development of this report form.

devised by Hanfmann and Kasanin⁷ will be described. All of these are based upon the principle that emotional disturbances and personality disorders interfere with thinking processes, particularly with ability to form abstract concepts. The purposes of these tests are, therefore, to help the psychologist observe the subject's thought processes and to discover the extent to which maladjustment or mental illness has impaired his conscious thinking as revealed in efforts to solve problems requiring the formation of concepts.

In particular these tests are intended to evaluate the subject's ability to deal with objects and situations on the abstract or conceptual level as compared with the concrete. Ability to form concepts implies conscious reasoning at the abstract level; that is, transcending the immediate specific sensory situation, abstracting the common property from particular instances, analyzing and synthesizing, shifting from one aspect to another, keeping in mind several aspects simultaneously, planning ideationally, and self-criticism. An individual's behavior at the concrete level, on the other hand, lacks these characteristics. The individual is then unreflective; he responds to the immediately given thing or situation as something unique; he does not perceive an object or situation as one instance of a general class or category.

The Goldstein-Scheerer Cube Test. This has been devised to determine whether the subject is able to copy colored designs with the use of cubes, each of which is colored thus: one side blue, one red, one yellow, one white, one blue-yellow and one white-red, the colors of the last two being divided diagonally. The twelve designs to be copied are, with one exception, those of the Kohs series.

According to the authors of this test, reproduction of the designs may be achieved through either a concrete or an abstract approach. In the former approach, the subject merely perceives the model as a whole and attempts, without analysis or reflection, to copy it. In the latter approach, the subject perceives the design reflectively and attempts to reproduce it while employing deliberate analytical reasoning directed toward a discovery of the relationships of the parts of the designs (presumably, the principles underlying its construction). If the subject fails to reproduce the design correctly on the first attempt, the examiner presents the same design in a graded series of

⁷ E. Hanfmann and J. Kasanin, "Conceptual Thinking in Schizophrenia," *Nervous and Mental Disease Monographs*, No. 67, 1942, p. 115.

modified forms, each of which is less difficult to apprehend than the preceding form. Each step is intended to facilitate the solution through: (1) enlargement to actual block-size; or (2) emphasis upon delineation of part relationships; or (3) actual use of block models. The subject is credited with having made a solution at the abstract level only if he succeeds with an original without the aid of the graded series of modified and simplified designs which are aids of a concrete nature. The authors state, "It would be absolutely erroneous to suppose that these aids help to initiate a process of abstraction in an individual who lacks the prerequisites for the act. These aids simply render that act [process of abstraction] unnecessary."⁸

Abnormal subjects, it was found, were unable to benefit from the presentation of graded concrete aids, whereas normal subjects did benefit. The one group were unable to learn; the other group did learn to solve the problem at the level of abstraction. For the normal group, the aids are said to be a means of learning to succeed at the abstract level on succeeding designs. For abnormal persons, the aids are concrete presentations without possibility of transfer value to subsequent situations. If the subject being tested cannot benefit from the modified and simplified aids, impairment of abstract behavior is indicated.

There are no objective criteria for scoring performance on this test. Experience in using it with normal and abnormal groups, however, is expected to provide the examiner with the qualitative criteria on the basis of which impairment or nonimpairment of abstract behavior will be shown. There are, however, *degrees* of impairment to be inferred from degrees of concreteness of behavior. Subjects who

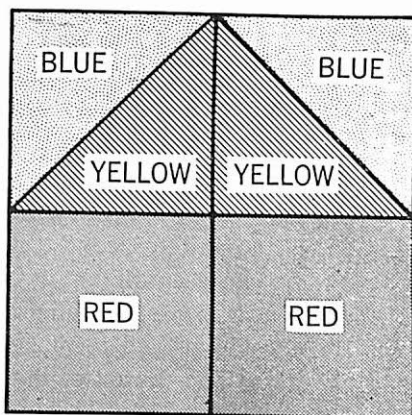


FIG. 16.1. From the Goldstein-Scheerer Cube Test. The subject is required to construct this pattern from a set of blocks, variously colored. Psychological Corporation. (By permission.)

⁸ Goldstein and Scheerer, *op. cit.*, p. 55.

need only the first of the facilitating series (enlargement of design) are regarded as less concrete in their behavior than those who need the subsequent steps (delineation; and finally, an actual block-model).

This cube test is considered by its authors to be suitable for studying impairment of abstract behavior in cases of mental deficiency, brain lesions, and dementia praecox.⁹

The Gelb-Goldstein Color Sorting Test. This is intended to examine the subject's behavior at the abstract or the concrete level on the basis of his ability or inability to sort colors according to definite color *concepts*. In one test, woolen skeins of different hue and tint are presented at random, there being twelve different shades of each color hue. The subject selects one and is then asked to pick out all others that go with it. In the second test, three skeins are presented; two are of the same hue but different in brightness and saturation; the third is of a different hue but the same in brightness as one of the first pair. The subject is expected to make a selection either according to hue or brightness. The third test involves the presentation of a series of six samples of the same color scale, from lightest to darkest red, and a second series of different hues but equivalent brightness. The subject is expected to perceive the common quality within each series. In the fourth test, the subject is asked to select all reds or all greens. In each of these four tests, the subject is asked to state the reasons for his groupings.

In the abstract approach, the subject treats the colored woolen skeins ". . . not as 'given things,' but as representative of their basic color hue." That is, the subject operating at the abstract level either selects *all* specimens of a given color without regard to differences in brightness, intensity, purity, etc., or he selects *various* colors of the same brightness, or same intensity, etc. In either case, the subject has discerned and used a concept based upon a common quality; and in so doing he has transcended the immediate sensory aspects of the situation. That is, he has used abstraction. In behavior at the concrete level, by contrast, the subject does not actually sort on the basis of a conceptual or categorical relationship; he merely employs a matching procedure.

The authors have found that abnormal persons with functional be-

⁹ A statement of evaluation will follow the presentation of the five tests by Goldstein and collaborators.

havior disturbances are incapable of assuming the abstract approach. In varying degrees, the abnormal subject appears to be unable to shift from concreteness of matching to the abstractness of categorization.

This color-sorting test is not scored numerically. The examiner, on the basis of his own interpretation, simply records the subject's performance as either concrete or abstract. The authors of the test, of course, provide specimens of both concrete and abstract behavior.

The G.G.W.S. Object-Sorting Test.¹⁰ This consists of a group of about thirty objects common to one's everyday experiences, there being one group for males and one for females.¹¹ Its purpose is "to determine whether the subject is able to sort a variety of simultaneously presented objects according to general concepts." Ability so to classify is evidence of the abstract approach; inability is evidence of the concrete approach. Classifications may be made on the basis of use (e.g., tools), color, form, materials, situational membership (e.g., implements for setting a dinner table), and pairings.¹²

The test employs several approaches. First, the subject is asked to pick out any object he cares to. After doing so, he is directed to select all other objects that he believes can be grouped with his initial selection. The examiner then himself selects some of the articles—one at a time—and again asks the subject to pick out others that belong with each initial item. Second, the subject is given all the disarranged articles and is told to place them in groups of his own devising. After the subject has made his own classification, the examiner urges the subject to find still other kinds of groupings. Third, the examiner himself makes several groupings representing still different categories of selection (e.g., color, form, toys) and asks the subject to state the concept-basis upon which the classifications have been made. In all of the foregoing approaches to conceptual classification, the subject is asked to explain what he has done and why he has done so, for the purpose of ascertaining whether he has consciously proceeded on a conceptual level of behavior.

Performance on this test, like the others in the group, is evaluated

¹⁰ By K. Goldstein, A. Gelb, E. Weigl, M. Scheerer.

¹¹ For example: male's includes toy hammer, plate, pipe, matches, knife, fork, apple, screw-driver, nails, candle, sugar; female's includes pencils, letter-opener, knife, fork, napkin ring, scissors.

¹² The inclusion of "pairings" as a form of classifying is of very doubtful validity; for pairing of two identical articles is nothing more than matching.

entirely on a qualitative basis, no norms or scale of performance being provided. The authors do report, however, that children of eight and nine years have been able satisfactorily to perform these tasks of classification. Brain-injured persons and some types of schizophrenics could not themselves make the abstract classifications, thereby revealing an impairment of mental functioning. Mentally deficient, also, were unable to make the required classifications; but their inability in this respect is quite within expectation, since very limited and arrested ability to deal with the abstract is a characteristic of that group.

Rapaport and his colleagues have devised a scheme for scoring this test.¹³ Their work indicates that it is possible, by means of scores, at least to facilitate an over-all survey of the subject's responses. No total score is derived; each of the several approaches (or parts) of the test is scored separately. The scoring is based upon both the act of sorting and the verbalization; namely (1) adequacy of sortings and verbalizations: that is, the degree of deviation from the norm of that item; (2) conceptual level of verbalizations: that is, whether the subject's definition is on the abstract, functional, or concrete level; (3) concept span: that is, the inclusion of all objects that belong in a well-defined category (the selections being neither too restricted nor too inclusive), utilizing a balance of the inductive and deductive processes. Although this plan of scoring does not have the degree of objectivity of standardized tests of intelligence and specific aptitude, it is, nevertheless, a distinct contribution in that it attempts to formalize evaluation of performance by selecting and defining criteria.

The Weigl-Goldstein-Scheerer Color Form Sorting Test. This is basically the same as that of the other Goldstein tests. Here the subject is required to sort a variety of differently colored geometric figures according to color or form. There are four equilateral triangles, four squares, and four circles. In each of these forms, one is red, one yellow, one green, one blue, while the reverse sides are all white. Conceptual behavior is indicated by the subject's ability to arrange the figures on the basis of color or form and to verbalize his act to confirm its conceptual character. It is sometimes difficult or impossible to judge from the *act* itself whether an individual's sorting is at

¹³ *Diagnostic Psychological Testing*, Vol. 1, pp. 401 ff. See also F. Boyd, "A Provisional Quantitative Scoring with Preliminary Norms for the Goldstein-Scheerer Cube Test," *Journal of Clinical Psychology*, Vol. 5, 1949, pp. 148-153.

the concrete or abstract level. Verbalization of the act is, therefore, essential.

Concrete perception and behavior are generally characterized by certain definite aspects, such as the following: a strong tendency to build patterns following structural lines; an inability to *account* for a grouping or to grasp the *meaning* of sorting; dependence upon sensory aspects and an inability to shift voluntarily from one sensory impression to another; inability to generalize from one performance to another.

Inability to approach the test at the abstract level is indicative of disturbance of cortical functioning. Here again, the examiner's evaluation of a subject's performance is qualitative, there being no scoring norms.

The Goldstein-Scheerer Stick Test. This is intended to examine the subject's ability to: (1) *copy* relatively simple geometric figures composed of sticks which are 3.5 inches and 5.5 inches in length; and (2) *reproduce* these same figures from memory, after exposure of 5 to 30 seconds. The sequence in which the thirty-four figures are presented represents something of a scale in terms of number of sticks involved and increasing complexity of the model.

As in the other Goldstein tests, the subject is asked to explain his performance, in order that the examiner may evaluate it as being concrete or abstract. If the subject's performance and report indicate that he has perceived and responded to "purely directional features in space"; if the subject perceives the figures as "mere configurations of spatial direction in a detached, purified sense"; if to the subject the figures "imply both space and direction as entities in themselves," then the performance is on the abstract level. For, in that event, presumably, the figures bear "no direct reference to a tangible life situation." Apparently, at the abstract level, the subject is expected to survey the pattern, and analyze it into its parts and part-relationships (space and direction).

Reference to the stimulus-figure as being associated with or representative of an object in actual experience constitutes a concrete response. Thus, to reproduce the stimulus-figure F correctly and call it an "F," or the figure A and call it a "roof," is to perform at the concrete level.

In this stick test, more than in the others of Goldstein, the exam-

iner is required to exercise his psychological insights in often making a subtle distinction as between a concrete or an abstract response.

The authors have found this test to be "particularly suited for cases with greater mental defectiveness or deterioration."

Evaluation of the Goldstein Tests. The authors of this series of tests do not present the type of statistical information upon which evaluations are usually based. In fact, there has been no attempt on their part to standardize their tests with respect to the usual criteria of validity and reliability. Their emphasis is placed entirely upon qualitative evaluation of the subject's responses to the problem situations as an aid in the diagnosis of mental impairment or of mental arrest. There is no doubt, however, that the clinical value of the Goldstein tests would be enhanced if quantified or scaled ratings could be obtained, not necessarily to obtain percentile or standard scores, or the like, but to facilitate an overall estimate of an individual's responses and to increase the objectivity of the tests' interpretation. At the present time, these devices appear to be valuable in the hands of an experienced psychologist for the purpose of detecting cases of marked intellectual impairment or arrest through the presentation of problems involving tasks of abstraction which are rather simple for persons of average ability who are functioning normally. In fact, some parts of this series of tests are at a difficulty level where most children of seven or eight years, and in other parts at nine years, can achieve a solution at the "abstract" level. Hence, in most instances, the failure of an adult to offer a solution, at the abstract level, with not too much difficulty, may well be regarded as a significant symptom of loss of intellectual level and efficiency.

The Hanfmann-Kasanin Test. This test consists of twenty-two blocks, each being in one of five colors, six shapes, two heights, and two widths. The problem for the subject is to discern how the blocks may be divided into four categories; namely: tall-wide, flat-wide, tall-narrow, and flat-narrow figures.

The subject is shown the entire set of blocks, randomly arranged. The examiner selects one as a sample and directs the examinee to pick out all others that are of the same kind. It is obvious that the subject might at first make his selection according to color, shape, or size. Each type of block has a nonsense name on the bottom, concealed from the subject (e.g., *bik* for flat-wide). After each grouping, the examiner shows the subject one of the wrongly selected forms

by revealing that it carries a different name. The procedure is to continue this kind of aid until the subject discovers the predetermined classification, if possible.

Performance is analyzed and scored in respect to interpretation of the task, nature of the attempts at solution, and discovery of the correct solution. In each of these, three levels of performance are distinguished: the primitive, the intermediate, and the conceptual, these being scored 1, 2, and 3, respectively. The principle of the scoring method may be generalized by saying that it is based upon the nature of the subject's approach to the problem, his ability to conceptualize, and his ability to verbalize his performance.

Evaluation. While in the development of this test its authors were concerned principally with schizophrenics and their clinical differentiation, Rapaport found it clinically useful in obtaining evidence of the subject's modes of thinking and response to difficult and frustrating problem situations, rather than for diagnostic differentiation.¹⁴ For example, different individuals may show the following modes of response in varying combinations and degrees: *fluidity* (lack of direction); *flexibility* (varying the approach, but keeping the end in view); *rigidity* (resistance to modification of behavior); *persistence* (continuity of behavior). These qualitative descriptions, therefore, obtained with the Hanfmann-Kasanin test are of the kind that are valuable as a supplement to numerical ratings in obtaining a more nearly complete description of a person's mental functioning.

Although the authors themselves used a scoring plan, the weights of the assigned values have not been experimentally determined. They are, rather, arbitrarily assigned scale-values; they may, therefore, be more properly designated as numerical indicators or identifications. As such they are useful in obtaining an over-all evaluation of an individual's performance on the test.

The problem-situations presented in the Hanfmann-Kasanin test are at a more difficult and higher level of abstraction than those in the Goldstein series. Since these problems make greater demands upon subjects of higher mental levels, they may reveal deterioration in conceptual thinking that would still be unapparent in the subject's long-established responses for meeting familiar situations and for dealing with familiar problems. The Hanfmann-Kasanin test, like the Goldstein series, provides a means of observing behavior in a controlled

¹⁴ Rapaport, *op. cit.*, Chapter 4.

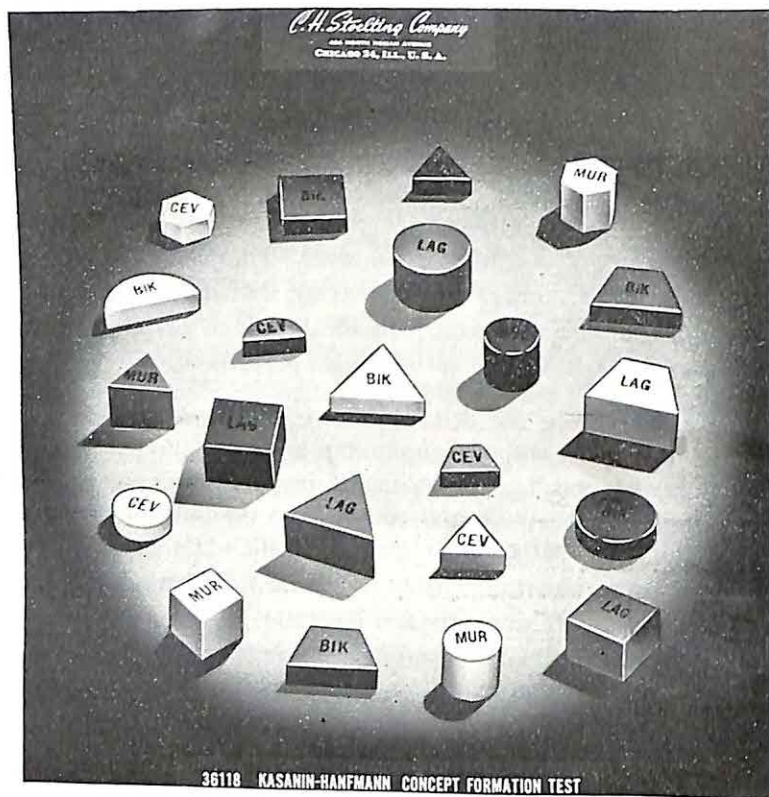


FIG. 16.2 Hanfmann-Kasanin test blocks. (C. H. Stoeckling Co. By permission.)

situation and of obtaining information of some significance to add to other psychological data and information.

THE HUNT-MINNESOTA TEST FOR ORGANIC BRAIN DAMAGE¹⁵

This test has been devised as an aid in clinical detection of organic brain damage, in the case of individuals who are sixteen

¹⁵ By H. F. Hunt. Published by University of Minnesota Press, 1943. See P. E. Meehl and M. Jeffery, "The Hunt-Minnesota Test for Organic Brain Damage in Cases of Functional Depression," *Journal of Applied Psychology*, Vol. 30, 1946, pp. 276-287; H. Juckem and J. A. Wold, "A Study of the Hunt-Minnesota Test for Organic Brain Damage at the Upper Levels of Vocabulary," *Journal of Consulting Psychology*, Vol. 12, 1948, pp. 53-57.

years of age or older.¹⁶ The instrument consists of three major divisions: the vocabulary test of the 1937 Stanford-Binet, which has been found relatively insensitive to brain damage; six memory and recall tests, which are considered to be sensitive to brain damage; and nine interpolated tests which serve as "validity indicators."

The six deterioration tests involve memorizing and retention of paired designs (presented visually, of course), and paired words presented orally. Both types are used to test immediate and delayed recall. A series of paired designs is exposed, without interruption, for six seconds each, after which the subject is shown one of each pair, in sequence, and is required to identify the design associated with it (immediate recall). In the word test, a series of ten pairs of words is read; after this is done, the first word of each pair is given singly, and it is the subject's task to name its paired word (immediate recall). There are three tests of designs and three of words.

The vocabulary test, as in the case of other clinical instruments (e.g., the Babcock) is regarded as being relatively stable, as holding up against deterioration. The vocabulary score, which is the number of words correctly defined according to the Stanford-Binet standards, is taken as the base for the determination of deterioration score, from which presence or absence of brain damage is inferred.

The interpolated tests consist of the following: information; naming the months of the year; counting from 1 to 20; counting from 3 to 30 by 3's; tapping on the table every time the number 3 is read in a long series of digits (attention test); counting backwards from 25 to 1; repeating digits backwards; naming the months in reversed order; subtraction of 3's from 79 to 1. These items are included for the following given reasons: as "validity indicators," since persons unable to perform these are too uncooperative, or too disturbed, or too deteriorated to be tested; hence testing them will not yield valid results. Critical scores are given for each of the interpolated tests, these having been reached or exceeded by ninety percent of the brain-damaged persons used in standardizing the test. The test's au-

¹⁶ Subsequent to publication of the test, Hunt reported that its maximal validity is for persons between the ages of 20 and 55 years, whose vocabulary scores are not at either one of the extremes of the distribution (less than 12 words nor more than 32). See H. F. Hunt, "A Note on the Clinical Use of the Hunt-Minnesota Test for Organic Brain Damage," *Journal of Applied Psychology*, Vol. 28, 1944, pp. 175-178.

thor reports that individuals whose scores fall below the critical levels in three or more of the interpolated items cannot be validly tested.

The interpolations also provide the means of filling a time interval after which retention of the same paired designs and paired words (delayed recall), given at the beginning of the examination, is tested.

Evaluation. This test was standardized upon a small number of patients (only 33) in several state hospitals who had been diagnosed as suffering from organic cerebral damage, excluding congenital conditions, birth injuries, and childhood brain injury. In age they ranged from 16 to 70 years. The control group consisted of 41 cases in state institutions, neuropsychiatric wards, or war veterans' hospitals, but who were *not* diagnosed as cases of organic brain damage.

When twenty-five cases from each group (brain-damaged and control) were equated on the basis of only their vocabulary scores, striking differences were found on the deterioration tests, and much smaller differences on the interpolated items. On the deterioration tests, using raw scores, the critical ratios of the differences between the two groups was 6.8 or more, whereas on the interpolated tests none of the critical ratios was greater than 3.8. The total overlapping of deterioration scores of the two groups was 50 percent, while total overlapping of interpolated scores was 90 percent. These data indicate that the former tests are much more discriminative between the two groups than the latter, and presumably, therefore, much more sensitive to effects of deterioration. The fact, however, that there is a 50 percent total overlap indicates that the test results must be viewed not in isolation, but in conjunction with other evidence in each case.

The correlation between deterioration scores and vocabulary scores was found to be $-.51$; between age and vocabulary, $.07$; between age and deterioration score, $-.37$. Multiple correlation of deterioration score with age and vocabulary was $-.65$.

This instrument, like all others in this difficult area of psychological testing, needs to be more clearly validated by further application and experimentation; for it appears that in some groups it does not differentiate as well as it should, too large a percentage of normal subjects having got scores which would suggest brain damage.¹⁷ It is well, therefore, to view this test at present—as suggested by its author

¹⁷ R. F. Malamud, "Validity of the Hunt-Minnesota Test for Organic Brain Damage." *Journal of Applied Psychology*, Vol. 30, 1946, pp. 271-275.

—as additional evidence offered in support of or contrary to a clinical suspicion of organic brain damage. At the same time, the more extreme positive deterioration scores have more diagnostic value.

THE BENDER VISUAL-MOTOR GESTALT TEST¹⁸

This test consists of nine figures which are characterized chiefly by their patterning, or configuration (i.e., their *gestalt*). The subject is simply instructed to copy each figure, without time limits, while it is before him. The test, clearly, is not one of visual memory or imagery; it is, rather, one of perception and of visual-motor functioning.

The figures used were devised by Max Wertheimer—one of the founders of the Gestalt school of psychology—in his experimental work on perception. The underlying principle utilized in this Bender test, as expounded by Wertheimer and others, is that organized wholes (structured units) are the primary forms of perception in human beings. Disturbances of perception (loss of integrative perception), therefore, might be psychopathological manifestations. Perceptual behavior is regarded, in the test, as involving sensory reception of the figures, interpretation at the central levels of the nervous system, and motor performance (drawing). This total process of perception and reproduction can be distorted by neural injury, by emotional maladjustment in the perceiving individ-

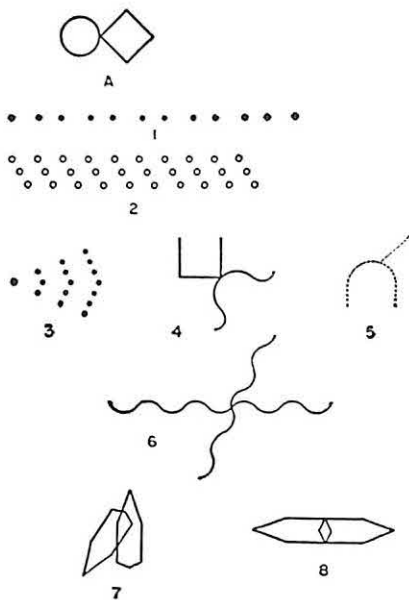


FIG. 16.3. The Bender Visual-Motor Gestalt Test. A Visual-Motor Gestalt Test and Its Clinical Use. Research Monograph No. 3, American Orthopsychiatric Association, 1938.

¹⁸ L. Bender, *A Visual-Motor Gestalt Test and Its Clinical Use*, Research Monograph No. 3, American Orthopsychiatric Association, 1938; F. Y. Billingslea, *The Bender-Gestalt: An Objective Scoring Method and Validating Data*, Clinical Psychology Monographs, No. 1, 1948.

ual, and by variations in the level of intellectual performance. Hence Bender explored the possibilities of this test of perception by investigating the "gestalt functions" in cases of aphasia, organic brain disease, schizophrenia, manic depressive psychoses, mental defectives, malingerers, psychoneurotics, and normal children.

Uniform directions for administering the test and scoring or analyzing the results have not yet been evolved. M. L. Hutt, who used this test extensively in the armed forces, has suggested that the drawings be evaluated for the following principal aspects:

Arrangement of the drawings: order (methodical, logical, irregular, confused); cohesion (expansive or compressive style); use of margin as a guide.

Modifications in size: reduction or expansion.

Use of white space: space left blank between drawings.

Modifications of the Gestalt: elaboration, distortion, destruction.

Motor incoordination: poor motor control representing physiological tensions or poor muscle tonicity.

Bender herself offers a different evaluation scheme, the essentials of which are:

Movements used in making drawings: e.g., speed, rhythm, perseveration.

Form of the drawing: e.g., outline, arrangement, spatial orientation, form differentiation, size, omissions.

In either case, the examiner also observes the behavior of the subject during the testing process, noting the methods used, verbalizations, and attitudes.

Subsequent to both of the foregoing analytic schemes, Pascal and Suttell provided a standardized and quantitative system of scoring the reproductions of adults.¹⁹ Essentially the scoring procedure was arrived at thus: (1) the reproductions of psychiatric patients were compared to those of normal persons; (2) the drawings of abnormal persons tended to deviate from the originals more than did those of normal subjects; (3) deviations (differences between originals and reproductions) that discriminated between normal and psychiatric subjects were isolated; (4) a deviation was retained if item analysis showed that it discriminated significantly between the two groups, or if

¹⁹ G. R. Pascal and B. J. Suttell, *The Bender-Gestalt Test*, New York: Grune and Stratton, 1951.

it occurred in the reproductions of abnormals but "practically never" in those of normals; (5) deviations were weighted according to discriminative value between the two groups; (6) "score reliability" was determined ($r = .90$; $N = 120$); (7) norms were found for a non-patient population of 474 subjects (271 males; 203 females) varying in age from 15 to 50 years, most of whom were attending evening classes at the high-school or college level.

After the scoring method had been evolved and the norms determined, validity was studied by: (1) "blind" matching of scores with group classification (normal, neurotic, psychotic); and (2) prediction

TABLE 47
Bender-Gestalt Test Means and Ranges of Scores

	Mean Scores *	Middle 60%	Range
Normals	50	47-60	32-79
Neurotics	68	53-80	32-139
Psychotics	81	65-100	40-155

* Values are in terms of Z scores; mean equals 50, standard deviation equals 10. The higher scores are the more unfavorable.
(Based upon data in Pascal and Suttell, *op. cit.*, pp. 30-31.)

of improbability of patients receiving therapy, tested on admission. The test was fairly effective in distinguishing between normals and psychotics and between normals and neurotics; but it discriminated only slightly between psychotics and neurotics. As between patients who were described, upon discharge, as "improved" or "unimproved," under (2), the mean scores for the two groups showed significant differences. The results under both (1) and (2) were sufficiently discriminating to encourage the use of the Bender-Gestalt as one possible source of significant evidence in the diagnosis of normal and abnormal personalities.

As in all such situations, it must be noted that while the mean scores and standard deviations for the *groups* are significantly different, there is still significant *overlapping* of scores of the three groups (normals, neurotics, psychotics). Overlapping of scores within different groups, classified for various purposes, is the usual psychological phenomenon. In dealing with an individual, therefore, in respect to a particular trait or function, it is always necessary to consider the possibility that he might deviate from the central tendency of his group. Specifically, the

scores on the Bender-Gestalt will illustrate this point, as shown in Table 47.

EVALUATION OF TESTS OF IMPAIRMENT

Psychological tests in this area are based upon the principle that old, well-established habits and modes of behaving (such as word knowledge) show *relatively* little loss, whereas new learning, newly acquired associations, performing new tasks and solving new types of problems are impaired in cases of brain damage and other forms of mental disturbance.

As a group, these psychological tests have not been adequately standardized, so far as norms, scoring, and interpretation are concerned. It must be recognized, however, that standardizing tests for the extremely deviant and often uncooperative populations for whom these tests are intended is an extraordinarily difficult task. Since these tests are intended for adults primarily, their standardization and interpretation are further complicated by the fact that allowance should be made for normally expected loss at advanced ages; and individual differences in cultural, educational, and occupational backgrounds must be considered in evaluating a subject's performance. Validation is made the more difficult, too, since one generally used criterion of validity is psychiatric diagnoses and classifications; and these are themselves sufficiently inconsistent and lacking in reliability as to introduce an important source of error in the validation process.

These tests do not, as yet, provide a self-sufficient method for measuring mental deterioration, except in the more marked cases. However, clinicians who have used them are in substantial agreement that they have value in that they provide opportunities for observation of mental operations under controlled conditions, in which prescribed materials are used. In such situations, an experienced psychologist is able to make important qualitative observations, in addition to deriving, at times, quantitative values for the subject's performance. The particular qualitative observations that can be made will depend upon the content and technique of the test. The observations on schizophrenics quoted on page 434 provide relevant illustrations. Other qualitative observations might include descriptions of the subject's thought processes, estimates of levels of abstraction or concreteness (as in the Goldstein and Hanfmann series), bizarre responses, degree of self-criticism, fluctuations of attention, degrees of rigidity or flexi-

bility of thought processes, level of immediate and delayed recall compared with recall of remotely learned materials, and evaluations of the subject's performance in the light of his former educational and occupational status. To make these observations, of course, requires a background of experience with a sufficiently large and varied number of subjects, including persons within the normal range of adjustment and performance.

PERSONALITY RATING SCALES

DEFINITION OF PERSONALITY

This term has been variously defined because personalities are complex and inclusive of all traits; hence, there is much room for differences in comprehensiveness of the definition. Those definitions, however, which include only the *social* value of an individual to other members of his group (that is, more or less superficial attractiveness or reputation) must be rejected as inadequate because they are concerned only with overt behavior, while they ignore the *inner* aspects of the personality: the perceptions, feelings, reactions, attitudes, values, prejudices which are the *basis* of one's behavior. These social-value definitions, in other words, are concerned only with what a person does and the impression made by him upon others in his social groups. Such impressions and evaluations are, of course, important in an individual's life; and they are evaluated by means of *rating scales*, which will be presented in this chapter. These definitions do not, however, take into account what *are* the basic traits of a person, aside from what he *actually does*, which might or might not reveal the covert aspects of his personality. As a matter of fact, some parts of the *personality inventories*, as distinguished from rating scales, are intended to identify these covert traits in order to provide the psychologist with a basis for a fuller understanding of an individual's behavior. Whether they succeed in doing so is a question to be discussed. Also, *projective methods*, more than any other type of instrument, are intended to reveal the covert, subtler

aspects of personality and behavior. These methods will be dealt with in the following chapters.

A definition which commends itself is the following: a personality is the product of the dynamic and unique organization within the individual of psychobiological structures, or systems, and their interaction with the environment. It is these two aspects—uniqueness of the structured organism and the characteristics of his environment—that determine the individual's particular adjustments to his surroundings.¹ A personality is the individuality that emerges from interaction between a psychobiological organism and the world in which he has developed and lives.

Personality is *described* in terms of an individual's behavior—his actions, postures, words, and attitudes and opinions regarding his external world. But personality is described, also, in terms of the individual's covert *feelings* about his external world; feelings which may not be apparent or discernible in his overt behavior. Furthermore, it is described in terms of one's feelings about *himself*.

One's actual feelings about his external world and about himself may be at the conscious, preconscious, or unconscious level. The same is true regarding the consciousness levels of the *reasons* for these same feelings. In other words, a person may know why he feels as he does (conscious); or the reasons for his feelings may be somewhat below (figuratively) the level of awareness, but such as to come to awareness with relatively little effort under appropriate stimulation (preconscious). Or the factors may be so deeply submerged or blurred that they can be brought to the level of awareness only with difficulty, if at all (unconscious). This being the case, then, it is desirable to have tests of personality which can probe the various aspects of personality and the levels of consciousness as well.

Several aspects of the foregoing definition need a little more explanation before the instruments themselves are presented.

By *dynamic organization* psychologists mean that personality traits do not exist independently or act in isolation. They are interrelated,

¹ For differing emphases in defining personality, see G. W. Allport, *Personality, A Psychological Interpretation*, New York: Henry Holt, 1937, p. 48; and Gardner Murphy, *Personality, A Biosocial Approach to Origins and Structure*, New York: Harper, 1947, Chapter 1. These two books are, to the present writer, the outstanding treatises of recent years on the subject.

interacting in an organized and coherent manner. They may, like any other organized system, be in process of change and evolution. *Disorganization* of traits and behavior results in "abnormality." It is in respect to the dynamic organization of traits that one of the great difficulties of personality testing is encountered, for it is much simpler to attempt the construction of an inventory that will give an indication of an individual's tendencies to introversion, or sociability, or self-confidence, or ascendancy, etc., than it is to measure or test the person as a whole. The reader will note, later, that actually the available inventories and rating scales deal only with smaller or larger segments of the personality, a number of which may be portrayed on a psychological profile; but to be meaningful these must somehow be organized into a meaningful whole by the psychologist who studies the individual case.

The term *psychobiological structures* connotes motives, habits, traits, attitudes, feelings, values, ways of thinking and acting. The word "psychobiological" is used to indicate that personality and its component integrals are neither exclusively mental nor exclusively neurological. Rather, they involve psychological processes and functioning (mind) together with their biological correlates (body).

Interaction with the environment is made explicit in order to emphasize that an individual's personality does not merely grow from within. It is the product of the interaction between himself as a developing organism having certain psychological and biological needs, on the one hand, and, on the other, his environment which has nurtured, influenced, directed, satisfied, or in varying degrees failed to satisfy those needs.

RATING SCALES: MAJOR ASPECTS

This type of device is useful chiefly to learn what impression an individual has made, in respect to some specified traits or attitudes, upon persons with whom he has come in contact. To a lesser extent it is also used for self-ratings. It is a device with which "social value" is rated in certain specified areas; it reflects the impression the subject has made upon the persons who do the rating. The rating of one person by others is among the oldest of practices, the present psychological tools being refinements upon the common practice of providing letters and oral recommendations.

For the evaluation of an individual, rating scales are submitted to

teachers, counselors, employers, colleagues, parents, and others who have had sufficient contact with the person in question to have formed an opinion based upon evidence. Usually, of course, ratings of a particular person are obtained from more than one judge; for validity of ratings is thereby increased, inasmuch as subjectivity of judgment is decreased through the balancing of errors and bias.

Rating scales may be devised for a great variety of personality traits: tact, generosity, leadership, cooperativeness, resourcefulness, punctuality, industriousness, honesty, emotional control, study habits, personal attractiveness, and many others, the number of possibilities being virtually unlimited. Each scale usually includes traits to be rated individually, the specific ones depending upon the purposes for which the scale is intended. The terms used to designate each of the several traits being rated are often vague in themselves and may have different meanings for different judges. In order to minimize this problem of semantics and to make the rating scales more useful than they would otherwise be, it is necessary to observe certain established principles. The following, then, are the major aspects to be considered in their construction and use.²

The traits must be clearly defined. This is essential so that they may be clearly and uniformly understood by all judges. This end may be achieved by giving explanations, synonyms, or specific instances as behavioral illustrations.

The degrees of the trait must be clearly defined. Each trait is rated on a scale generally having four or more intervals, five and seven being the most frequent. A larger number of intervals requires refinements of distinctions and ratings which are not often possible. Each step on the scale of each trait must be clarified in much the same way as the trait definitions themselves.

The scale may be of the scoring or ranking type. These are the two basic types, although there are variations. In the first type, the subject is rated at a point, or level, on the scale without direct reference to or comparison with other persons in his group (his classroom, his fellow workers, his fellow club members, and the like). Each point, or level, on the scale carries a specified score. On a scale of five, for example, the "average" person may be scored 0; in which case the deviants would be scored -1, -2, +1, or +2. Or the scores

² Cf. Allport, *op. cit.*, Chapter 16.

might all be positive, the lowest rating being 1, the average 3, and the highest 5.

A common variant of the scoring scale is the *graphic rating scale*. The several levels, or degrees, of the trait are defined and placed at points along a horizontal line. The judge places a mark anywhere he chooses on this line, between the two extremes. Although a graphic scale permits, theoretically, scoring at a very large number of points, such refinement and spurious accuracy are not warranted. The investigator or compiler of information will, therefore, convert each rating within a given range according to a predetermined numerical scheme.

The following are examples of the numerical rating device.

How emotional is the parent's behavior where the child is concerned?

- _____ Constantly gives vent to unbridled emotion in response to child's behavior.
- _____ Controlled largely by emotion rather than by reason in dealing with child.
- _____ Emotion freely expressed, but actual practice is seldom disorganized.
- _____ Usually maintains calm, objective behavior toward child, even in face of trying situations.
- _____ Never shows any sign of disorganization toward child.

Does he get others to do what he wants done?

- _____ Display marked ability to lead his fellows
- _____ Sometimes leads in important affairs
- _____ Sometimes leads in minor affairs
- _____ Lets others take the lead
- _____ Probably unable to lead his fellows
- _____ No opportunity to observe

Each rater checks what he believes to be the correct description. The investigator then will convert the check marks into scores. In this instance, the third item would be the average and would be scored zero; the first and second would be -2 and -1, respectively; the fourth and fifth, +1 and +2, respectively.

The graphic type of item may be illustrated by the following, remembering the necessity of definition of the trait and clarification of levels.

Quality of work

Of doubtful satisfaction.	Not quite up to standard.	Satisfactory.	Superior to general run.	Exceptionally high.
---------------------------	---------------------------	---------------	--------------------------	---------------------

Attitude toward others

Quarrelsome, uncooperative, upsets morale.	At times difficult to work with.	Ordinarily tactful, cooperative, and self-controlled.	Always congenial and cooperative.	Unusually strong factor in co-operation and group morale.
--	----------------------------------	---	-----------------------------------	---

The *ranking* scale is used with persons who are associated within a single group and who are to be rated relative to one another. The judge arranges the names in serial order with regard to each one's status in a specified trait. Usually, the judge is instructed first to select the individuals to be ranked highest, lowest, and average, and then to place the others in relation to these three. Since intervals between successive individuals are not equal, and since it is impossible by this method to determine the sizes of the intervals, arithmetical and statistical computations are not warranted and should not be attempted. The ranking scale simply provides a method for use with a single group of subjects when, for any valid reason, intragroup comparisons are desired.

Reliability depends upon extent of variation of judges' ratings. Judges rating an individual on a specified trait will not always agree as to his score or rank. It is customary, therefore, to take the mean of all the judgments as representing the nearest approximation to the true rating. If this method of averaging is to be meaningful, however, it is essential that the variation of the judges' ratings shall be small, thus indicating reasonably close agreement. A large variation, on the other hand, would indicate unreliability due to lack of clarity regarding the trait being evaluated, contradictory or unstable behavior of the subject, or undependability of some judges. An average of the judges' ratings, without regard to their variation might be misleading or even absurd. For example, if, on a seven-point scale, two judges rated an individual at -3 , two at $+3$, and two at zero,

the mean rating would be zero (or average level), whereas the probability is that he is not average at all in view of the wide disparity of judgments. Reliability of ratings is usually dependent upon having a sufficient number of qualified judges, five to seven being the number frequently recommended. The degree of agreement required before a set of judgments may be regarded as sufficiently reliable is, to some extent, arbitrarily determined; but to be statistically reliable, agreement among judges should be three or four times as great, at least, as that obtained by chance.

At times it is possible to find out, by inspection of all the ratings, which *judges* seem to be most dependable by noting the extent to which each of their ratings approximates the mean of all ratings of a trait. Similarly by inspection, it is possible to identify those subjects who have been most reliably rated, in terms of agreement among the judges. Thus, certain rating scales can be useful in some *individual* cases, whereas the *group* trends and evidence might be less than satisfactory.

Methods of studying reliability of rating scales include the following, most commonly: repeating judgments after a time interval; correlation between ratings of two or more judges; and relationship between judges' ratings and self-ratings.³ The correlation coefficients thus found are in the neighborhood of .50 and .60—much lower than would be acceptable in the case of tests of general intelligence, specific aptitudes, or educational achievement. Occasionally, however, much higher reliability coefficients have been obtained, some being in the neighborhood of .85 to .90.

The relatively low reliabilities are not necessarily attributable to defects in the conception of a scale nor in the phrasing of the characteristics to be rated. The coefficients reflect, to a considerable extent, the differences between and the unreliabilities of the judges.

Determination of validity is difficult. There are few, if any, useful criteria of validity, in the usual sense, which can be employed in evaluating a rating scale. There are at present no measured, or measurable, behaviors against which these scales can be compared. There are no numerical values to be used as criteria, such as teachers' marks, achievement test scores, age differences in mean scores, etc. In fact, the subjective judgments sometimes used as validating criteria with

³ See J. P. Guilford, *Psychometric Methods*, New York: McGraw-Hill, 1936, Chapter 9.

other types of tests (e.g., intelligence, personality inventories) are the measures of *reliability* of rating scales, as indicated above. In other words, the validity of the rating scale is assumed, in actual practice, to rest upon the judges' understanding of the meanings of the traits being evaluated and their accuracy in rating them. The principal indication of validity of some rating scales is the fact that persons using them—guidance counselors, personnel officers, employers—find them helpful if the judges are carefully selected and if the ratings are conscientiously made.

Overt traits are more reliably rated than covert. Traits which can be rated upon objective activities, on the basis of actual past or present performances known to the judges, are most reliable. For example, general emotionality, social acceptability, aggression, fear, anxiety, and impulsion are rated with greater reliability than those dealing with a person's inner life and feelings about the self.⁴

Judges should be instructed. It has already been stated that the traits to be rated and the intervals on the scale should be clearly defined for guidance of the judges. In addition, it is necessary to instruct them with regard to other aspects affecting the reliability of their judgments: namely, each trait rating should be made independently of other ratings of the same person (avoidance of the "halo" effect); ratings should not rest upon inadequate acquaintance with the subject; experience and acquaintance with a broad enough variety of people to provide bases of judgment are desirable; sincere motivation to provide the most reliable ratings possible is necessary. Of the foregoing, the "halo" effect has been shown to be among the most serious causes of unreliability. One tends, for example, to over-rate a person in all respects if he likes him, or if he is a close acquaintance.

Judges should state their degree of certainty. With each rating, the judge should be asked to state his degree of certainty (e.g., very strong, strong, moderate).⁵ It has been demonstrated that judges are most confident and most in agreement on ratings at the extremes. This is understandable because extreme deviants are most distinguishable from others and are most readily characterized by the trait names. Thus, the terms cooperative-uncooperative, introverted-extroverted—

⁴ See R. Wolf and H. A. Murray, "An Experiment in Judging Personalities," *Journal of Psychology*, Vol. 3, 1937, pp. 345-365.

⁵ A similar technique is now often employed in polling public opinion.

these apply most forcibly and clearly to those individuals who are manifestly one or the other.

Some persons are more accurately rated than others. On the whole, extroverted individuals are more reliably judged than introverted. Quite understandably, the ratings of persons whose traits are characterized by overt behavior rather than by covertness and inner qualities will be based upon fuller, more representative, and better understood behavior samples. Also, it has been found that judges rate more reliably those persons who most resemble themselves because, it is believed, one can best empathize with people whose behaviors resemble his own.⁶

Reliability of trait estimates is affected by desirability or undesirability of the trait. In self-ratings there is a tendency for individuals to overrate themselves in respect to traits regarded as socially desirable. In rating other persons, especially friends, some judges may be similarly influenced even while trying to be conscientious. In fact, there is a general tendency toward generosity in ratings, rather than the reverse.

REPRESENTATIVE RATING SCALES

A few scales, representative of the more satisfactory ones, will be briefly described.

Haggerty-Olson-Wickman Rating Schedules.⁷ These are designed for the detection and study of behavior problems and problem tendencies in individuals from nursery-school through high-school levels. Schedule A is a behavior-problem record, enumerating fifteen types or sources of problems, such as speech difficulties and defiance of discipline. Each of the fifteen is rated from 1 to 4, depending upon frequency of occurrence. Schedule B is a graphic scale, consisting of thirty-five traits classified into four groups: intellectual, physical, emotional, and social. These traits are scored on a 5-point scale. The authors provide better than usual evidence of validity. They report a correlation of .76 with frequency of referral for reasons of discipline or other action by school principals, which may be symptomatic of a variety of adjustment difficulties. They report, also, that only ten per-

⁶ R. Wolf and H. A. Murray, *op. cit.*

⁷ World Book Co., 1930.

cent of "normal" children reach or exceed the median score of those referred to the psychological clinic.

The Vineland Social Maturity Scale.⁸ This scale is unique in having been constructed and standardized on the model of the Stanford-Binet scale. It is designed for use with individuals from infancy to the age of thirty years.

Unlike many other scales, this one is based upon a well-defined rationale and has been systematically constructed. Behavior items are grouped at age levels, as in the Stanford-Binet. The items represent progressive maturation and adjustment to the environment in the following categories: self-help, self-direction, locomotion, occupation, communication, and socialization. The following examples illustrate the several categories.

- Self-help: Reaches for nearby objects (age 0-1)
- Self-direction: Buys own clothing (age 15-18)
- Locomotion: Walks about room unattended (age 1-2)
- Occupation: Helps at little household tasks (age 3-4)
Systematizes own work (age 25+)
- Communication: Makes telephone calls (age 10-11)
- Socialization: Demands personal attention (age 0-1)
Advances general welfare (age 25+)

Items are scored after interviewing someone well acquainted with the subject, or the subject himself. A social age is then obtained; this is divided by chronological age, yielding a social quotient (S.Q.).

Although this social maturity scale shows a high correlation with intelligence test results (about .80), Doll maintains that it is distinct enough in content and in the functions rated to warrant its use in the study of an individual's general behavioral development, since social age provides a basis upon which to proceed in the care and training of an individual.

While the scale is intended for use with a normal population as well as with mentally deficient, it was first conceived as an aid in the diagnosis of feeble-mindedness. In the first instance, it was and still is intended to differentiate between mentally deficient individuals who are also socially inadequate, on the one hand, and, on the other hand,

⁸ By E. A. Doll. The Training School, Vineland, N. J., 1936. Also by Doll, "Preliminary Standardization of the Vineland Social Maturity Scale," *American Journal of Orthopsychiatry*, Vol. 6, 1936, pp. 283-293.

the mentally retarded who are able competently to conduct their personal and social lives.

The Vineland scale has had wide use in clinics for children and adolescents; because, in addition to the uses already indicated, it is a valuable device for interviewing and counseling both parents and children.

Progressive Education Association Behavior Description.⁹ This is intended for use by teachers. The traits included are the following: responsibility-dependability, creativeness and imagination, influence, inquiring mind, open-mindedness, power and habit of analysis, social concern, emotional responsiveness, seriousness of purpose, social adaptability, work habits, physical energy, assurance, self-reliance, and emotional control. This is a rather inclusive and varied list of traits which could be rated only by a judge who had had long and varied contact with the subject. It will be noted that the traits include complex intellectual activities as well as those more often found in personality rating scales.

The Fels Parent Behavior Scales.¹⁰ This device provides thirty rating scales, in as many aspects, for assessing parental behavior toward their children. Ratings are made by a qualified observer in respect to those aspects of the home environment that the rater has been able to observe and review in a systematic manner, by means of home visits and interviews with parents. Among the rated thirty aspects of child-parent relationships are the following: discord in home, sociability of family, child-centeredness of family, restrictiveness of regulations, readiness of criticism, rapport with child.

In clinical and research work, these scales have been found useful in providing a systematized approach to the detailed analysis of psychological environments in which children are developing. The scales provide a standardized method for describing and estimating the nature of the effect of a child's home upon his behavioral development.

Other Scales. The reader will have observed that the instruments listed above are all intended for school and clinical use. The reason is

⁹ Reports and Records Committee of the Progressive Education Association, 1938.

¹⁰ Fels Institute, Yellow Springs, Ohio, 1937-1949.

that in schools and clinics are to be found the better devices of this kind. There are, of course, innumerable rating scales that are being used in business and industrial personnel departments, social agencies, and counseling organizations;¹¹ but they are not different from those presented above, nor are they original in conception; and relatively few of them have been subjected to the scrutiny to which scales used by psychologists in schools and clinics have been subjected. The items that follow are from a pre-World War II rating scale prepared for and used in a specific industry; but the items in this scale are typical of those widely used in similar situations.

Relations with other supervisors:

- a. Often not satisfactory
- b. Sometimes not satisfactory
- c. Usually gets along well
- d. More satisfactory than average
- e. Exceptionally satisfactory

Knowledge of the characteristics and abilities of subordinates:

- a. Knowledge markedly limited
- b. Knowledge somewhat limited
- c. Knows employees fairly well
- d. Knowledge better than average
- e. Knowledge exceptional

Willingness to make difficult decisions:

- a. Often "passes the buck"
- b. Inclined to "pass the buck"
- c. Usually properly willing
- d. More willing than average
- e. Exceptionally willing

Resourcefulness in meeting difficulties:

- a. Often goes to pieces
- b. Easily discouraged by obstacles
- c. Usually meets the situation
- d. More resourceful than average
- e. Nearly always finds good way out

Ability to learn new work:

- a. Learns with difficulty
- b. Learns somewhat slowly
- c. Learns fairly easily
- d. Better than average
- e. Learns with exceptional ease and speed

¹¹ See, for example, W. D. Scott, *Personnel Management* (New York: McGraw-Hill, 1941), *passim*.

In the armed forces during World War II, rating scales were devised to assist in personnel evaluation and selection in a variety of situations.¹² These followed the usual principles and practices already discussed. They need not, therefore, be described here.

EVALUATION OF RATING SCALES

Rating scales are not tests; nor are they precise or objective measures; hence, their reliability cannot be so high as that found for other types of psychological instruments. Rating scales do provide a means of obtaining organized descriptions of behavioral traits from judges who have had ample opportunity to make the necessary observations. If the scale meets the specifications already presented in this chapter, these ratings will be based upon greater uniformity of trait-definition and trait-connotation than will purely individual ratings of traits defined by each judge independently and for himself.

By means of an organized scale it is possible to obtain ratings on specified traits which are considered essential or significant in the particular setting where the scale is being used. Completely independent, or "unstructured," ratings, by contrast, may fail to provide desired information.

While high reliability of ratings among judges of the same persons would simplify interpretations of results, moderate or even low reliability coefficients do not discredit a scale which is soundly conceived and developed. The reason for relatively low reliability coefficients may be one or a combination of the following two factors. (1) In spite of carefully defined traits and degrees thereof, ratings are always subject to the judges' biases, values, and standards of performance and behavior. Some degree of influence of these aspects of the judges' own personalities is inevitable. (2) Behavior of the person being rated may be and at times is to some extent variable in different types of situations. This is the case even though there is some degree of trait-consistency within a person. For example, an individual who is "self-confident" in one situation is more prone to be "self-confident" in another, whereas a person who is "withdrawn" is prone to behave that way in many settings. Yet that same individual is not equally self-confident or equally withdrawn on all occasions; and, in fact, there may be situations in which his characteristic mode of behavior is not

¹² See D. B. Stuit (editor), *Personnel Research and Test Development in the Bureau of Naval Personnel*, Princeton University Press, 1947, *passim*.

readily apparent. Variability of behavior is especially true of children and adolescents whose personality traits are still in process of formation. In interpreting ratings, it is necessary, therefore, to know the types of situations in which each judge made his observations.

The usual criteria and standards of validity are not applicable to rating scales. There is a matter of *internal* or face validity. The questions to be asked in regard to validity of a rating scale are these: Does it meet the specifications of a sound scale? Are the traits which are being rated by the scale significant in the setting or occupation for which the individual is being considered? If these two questions are answered satisfactorily, then the ultimate usefulness (i.e. predictive validity) of the scale will depend upon the soundness (reliability) of the judges' ratings.

PERSONALITY INVENTORIES

PURPOSES AND TYPES OF INVENTORIES

It has been estimated that there are in the neighborhood of five hundred personality tests and inventories. Obviously, there would be no point in attempting to describe all of them, nor even a large number; for they have much in common, and many of them are so inferior in conception and validation as to merit complete neglect. The inventories which are briefly presented in the following pages are among those that have been most widely used and are representative of the group as a whole. These criteria, however, by no means imply that they are wholly satisfactory instruments. They are more or less useful, within limits and in the hands of qualified psychologists. Their limitations will be indicated in the section on evaluation.

Rating scales, for the most part, are intended to reveal how other persons (the judges) respond to or have been impressed by the subject; that is, these scales provide evidence of the value placed upon the subject in certain group situations. Personality inventories, on the other hand, are self-rating questionnaires which deal not only with overt behavior (e.g., insisting on having one's own way, emotional expression, sympathetic acts) but also with the person's own feelings about himself and his environment, resulting from introspection (e.g., liking to be alone and living introvertly, need for praise, repression of desires, caution and worry). Insofar, then, as inventories actually get at aspects of personality which are beyond impressions made upon observers, and resulting reputation, they are the more valuable instruments.

Personality inventories may be classified into four types: (1) those that assess specified traits (e.g., ascendance, conservatism, self-confidence); (2) those that evaluate adjustment to several aspects of the environment (e.g., home, school, community); (3) those that classify into clinical groups (e.g., paranoiac, psychopathic personality); and (4) those that screen subjects into two or three groups (e.g., psychosomatic disorders *versus* normal). An example of the first of these is the Bernreuter; of the second, the California; of the third, the Minnesota Multiphasic; of the fourth, the Cornell Index. These inventories are among those described in this chapter.

The classification into four groups does not signify that the inventories in each group have nothing in common with the others. The differences between them are dependent upon purposes, organization and nature of content, and scoring categories. Fundamentally, nearly all personality inventories are based upon the principle that behavior and personality are, in part, manifestations of certain *traits*, and that the strength of traits can be measured.

A *trait* may be defined as a generalized mode of behavior or a form of readiness to respond with a marked degree of consistency to a set of situations which are functionally equivalent for the person. A trait is a form of adaptive or expressive behavior employed by the individual in situations that he perceives as having some equivalence. Thus, if a child readily volunteers information and opinions in school classrooms but is reticent in all other situations, his classroom behavior would be regarded as a "habit" rather than as a trait. However, if this child's classroom self-confidence extended into a variety of situations—that is, if it is generalized—his self-confidence would be designated as a trait. Also, if a person always votes for candidates of the most conservative party, this might be only a habit; but if conservatism is his characteristic mode of responding to a variety of situations (along a scale of conservatism-radicalism), then conservatism is one of his traits.

Thus, in personality inventories an effort is made to estimate the presence and strength of each specified trait through a number of items representing a variety of situations in which the individual's generalized mode of responding may be sampled. The traits selected for measurement in a particular inventory are those which are present in varying degrees, and can be compared, among the members of the population for whom the inventory is intended.

REPRESENTATIVE INVENTORIES

The Allport A-S Reaction Study.¹ The Allport is a self-rating inventory for college students and, in revised form, for business use, its purpose being to disclose the degree to which the individual tends to dominate others (ascendancy) or be dominated by them (submission). The inventory presents a variety of situations which are likely to confront a person in the normal course of daily living. For each situation, several possible answers are presented; the subject indicates the one which most adequately characterizes his behavior. For example:

If a student in class discussion makes a statement that you think erroneous, do you question it?

usually

occasionally

never

The Bell Adjustment Inventory.² This consists of questions intended to evaluate the subject's status in respect to: home (satisfaction or dissatisfaction with home life); health (extent of illness); social adjustment (extent of shyness, submissiveness, introversion); emotional adjustment (extent of depression, nervousness, ease of disturbance); occupational adjustment—of adults (satisfaction with work, associates, and conditions). There are two forms: one for students (grade 9 through college) and one for adults. The items are of the usual kind, to be answered as *yes*, *no*, or *?*. For example: "Are you troubled with shyness?" "Are you often sorry for the things you do?" "Do you daydream frequently?"

The Bell Inventory raises a problem which is common to all devices of this kind: namely, do the questions and the scores for each category actually represent separate and distinct aspects of behavior and adjustment? Are these aspects mutually exclusive? Some critics maintain they are not. They hold, on the contrary, that the same per-

¹ By G. W. and F. H. Allport. Houghton Mifflin Co., 1928-1939. Of the personality inventories, the Woodworth Personal Data Sheet (C. H. Stoelting Co.) is the oldest. Its items were based upon symptoms reported in studies of psychoneurotic cases. This inventory was adapted by Woodworth and Mathews for use with children (C. H. Stoelting Co.). These two devices have been largely superseded by more recent inventories.

² Stanford University Press, 1934-1938.

sonality variables influence adjustment in all situations and, therefore, that the more useful and significant inventories are those which probe the various psychological mechanisms such as hysteria, defense (e.g., rationalization and projection) and escape techniques (e.g., negativism, suppression), psychosomatic manifestations, etc. Other psychologists, while recognizing the instrument's inability to reveal the dynamics of behavior, nevertheless believe it is useful in placing the individual relative to a group in respect to the specified areas of behavior, and as a basis for further psychological interviewing. While the first criticism is warranted, the Bell inventory has found wide and justified use for the latter purpose.

The Bernreuter Personality Inventory.³ This instrument is a questionnaire intended for use in grades 9 to 16, and with adults. Although the items are not arranged into categories, they are scored for six traits: neurotic tendency, self-sufficiency, introversion-extroversion, dominance-submission, confidence, and sociability. The last two of these were added by J. C. Flanagan after factor analysis. The items themselves and the manner of answering (*yes, no, ?*) are not new. The method of scoring is, however, not typical; for each of the responses to each item is regarded as characteristic of several traits, the scores for each item being weighted on the basis of empirically or statistically determined differentiating power. There are thus six scoring scales, one for each of the specified traits.

This inventory, like others of its kind, has been criticized as having set up arbitrarily determined personality categories and as being inadequate for individual diagnosis. Its principal value is as an aid in identifying persons at the extremes of the scale, as an early step in their full psychological study. Several sample items follow: "Have you ever crossed the street to avoid meeting a person?" "Are you inclined to study the motives of people carefully?" "Do people ever come to you for advice?"

/ **The California Test of Personality.**⁴ Of the questionnaire type, this inventory is ". . . organized around the concept of life adjustment as

³ Stanford University Press, 1931-1938. See also J. C. Flanagan, *Factor Analysis in the Study of Personality* (Stanford: Stanford University Press, 1935).

⁴ By L. P. Thorpe, W. W. Clark, and E. W. Tiegs. California Test Bureau, Los Angeles, 1953.

a balance between personal and social adjustment." There are five scales: primary, elementary, intermediate, secondary, and adult. The questions, answered either "yes" or "no," are grouped under the following categories:

Personal adjustment: self-reliance, sense of personal worth, sense of personal freedom, feeling of belonging, withdrawing tendencies, nervous symptoms.

Social adjustment: social standards, social skills, anti-social tendencies, family relations, school relations, occupation relations (adult level only), community relations.

This broad two-fold division is consistent with a rather frequent practice of classifying adjustment difficulties into "personality" problems (personal adjustment) and "conduct" problems (social adjustment). But it should not be assumed that any particular focal maladjustment is restricted to only one or the other of these categories; for, in fact, the whole person and his environment are involved in behavioral difficulties or disorders. What these two major categories and their subdivisions do is assist in identifying some of the principal sources of an individual's problems. This inventory, like the Bell, provides an opportunity for responses that may be symptomatic of maladjustment and that can be valuable in subsequent psychological interview and treatment.

Several items from the intermediate inventory follow.

Do you keep on working even if the job is hard? (self-reliance)

Do you find that a good many people are mean? (sense of personal worth)

Is it hard for you to say nice things to people when they have done well? (social skills)

Do you often visit at the homes of your boy and girl friends in your neighborhood? (community relations)

Aspects of Personality.⁵ This is an inventory for use with pupils from grade 4 to grade 9. The items are specifically adapted to experiences of school children; and for the purpose of clear comprehension the vocabulary employed is at the fourth-grade level. (This aspect can be a disadvantage with older pupils who might respond unfavor-

⁵ By R. Pintner, J. J. Loftus, G. Forlano, and B. Alster. World Book Co., 1938. Items reproduced by special permission.

ably to its "babyish" quality.) The inventory is intended to reveal psychoneurotic tendencies, the sections purporting to rate ascendance-submission, introversion-extroversion, and emotionality. Several specimen items are: "I have a lot of nerve"; "I like to read before class"; "I feel tired most of the time." After each statement the subject indicates whether he feels the "same" or "different."

The authors of this instrument suggest that it be administered in class at the beginning of the school year ". . . in order to acquaint the teacher as soon as possible with the personality make-up of her children." This suggestion is not only unwise; it is dangerous. It implies that personalities can be studied by group or mass testing, that this inventory is comprehensive enough to portray pupils' personalities, that personalities are to be described and diagnosed in terms of test indexes, and that teachers are or should be equipped to make personality diagnoses and provide remedies for the maladjusted. None of these implications is warranted.

The Minnesota Personality Scale.⁶ This inventory, having one form for men and another for women, is constructed to rate the following aspects of personality: morale (belief in society's institutions and future possibilities); social adjustment (gregariousness and social maturity); family relations (parent-child relations); emotionality (degree of stability); economic conservatism (degree on a scale from conservatism to radicalism). The inventory is devised for use in the last two years of high school, with college students, and "in some adult cases." An aspect of this instrument infrequently found is the gradation of answers whereby the subject indicates the strength of his responses. Instead of the very common *yes, no, or ?*, the subject in this instance has five choices, such as: *strongly agree, agree, undecided, disagree, strongly disagree*; or *almost always, frequently, occasionally, rarely, almost never*. The score of each item is weighted from one to five, corresponding to the degree of intensity represented by the choice of answer.

Perhaps the only parts of this inventory needing special mention are the first and fifth: morale and economic conservatism. Both traits are unusual in personality questionnaires. In the first, whereas high scores are regarded as indicative of belief in society's institutions and

⁶ By J. G. Darley and W. J. McNamara. Psychological Corporation, 1941.

future possibilities, low scores "usually indicate cynicism or lack of hope in the future." Sample items on morale are: "No one cares much what happens to you." "Court decisions are almost always just." Two items on the scale of economic conservatism-radicalism are: "On the whole our economic system is just and wise." "Poverty is chiefly the result of injustice in the distribution of wealth."

The particular selection of the five personality aspects tested by the Minnesota inventory may appear to be a rather strange one. The authors explain their selection as being ". . . the result of . . . work on problems of personality measurement in a clinical personnel program" in the University of Minnesota. The personality aspects sampled with this instrument have been found valuable in identifying ". . . a substantial proportion of adjustment problems in a large scale student personnel program" after a number of traits and attitudes had been experimentally investigated.

The Minnesota Multiphasic Personality Inventory.⁷ We have here one of the most elaborate and ambitious instruments in this field. The authors state in their Manual that the inventory is ". . . designed ultimately to provide, in a single test, scores on all the more important phases of personality. The point of view determining the importance of a trait in this case is that of the clinical or personnel worker who wishes to assay those traits that are commonly characteristic of disabling psychological abnormality." It is intended for persons sixteen years of age or older who are able to read.

The inventory consists of 550 statements, each of which is printed on a separate card. The cards are sorted by the subject into three groups—*True*, *False*, or *Cannot Say*—depending upon whether he regards the statement as true of himself or not. The statements cover a wide range, including physical condition, morale, and social attitudes. The items have been classified under twenty-six headings—for example, general health, gastrointestinal system, family and marital, religious attitudes, affect (depressive and manic), delusions, phobias, masculinity-femininity interests. Various items of the inventory have been selected and grouped, thus far, to form separate scales for the scoring of nine personality traits. Other scales, based upon still different groupings of some of the 550 items, may also be developed. The

⁷ By S. R. Hathaway and J. C. McKinley. Published by The Psychological Corporation, 1943–1951. A booklet form is available.

nine are: hypochondriasis, depression, hysteria, psychopathic deviate, masculinity-femininity interest, paranoia, psychasthenia, schizophrenia, and hypomania. Since its original publication, scoring has been developed for a new scale from the same 550 items. This is called the social scale, to measure the tendency to withdraw from social contacts with others.

It is apparent from this list of scales that the Multiphasic inventory is concerned almost exclusively with the clinical problem of differential diagnosis. This is further indicated by the fact that the scales were developed by contrasting normal groups with clinical psychiatric cases. The chief criterion of validity was the prediction of clinical cases against the diagnoses of a hospital staff.

The 550 items that constitute the whole inventory are in themselves not unusual. The distinguishing characteristics of the inventory are: its comprehensiveness; its large number of scale groupings to diagnose clinical types; and four aspects of its scoring, namely a "validity score," a "lie score," a "question score," and a "K" score.

The first is based upon a group of items which serve ". . . as a check on the validity of the whole record. . . . If the [validity] score is high, the other scores are likely to be invalid either because the subject was careless or unable to comprehend the items, or because someone made extensive errors in entering the items on the record sheet. A low [validity] score is a reliable indication that the subject's responses were rational and pertinent." This score is obtained from sixty-four items which have been answered uniformly by about ninety percent of normal persons and by nearly as many miscellaneous abnormal subjects. It has been concluded, therefore, that a marked deviation from these uniform responses is an indication of invalidity of other responses, for reasons given above.

The "lie score" consists of fifteen items on which a high score probably indicates that the subject has answered them falsely in order to create a favorable impression and thereby place himself in a socially acceptable light. In general, these are statements about one's behavior which, if they apply to the subject (and they do apply to practically everyone), indicate that he is something less than perfect. Though a high "lie score" does not necessarily invalidate the other scores, it may indicate that some of them have been influenced similarly; thus the total result is open to question. There are, however, instances of individuals whose "lie scores" are not indicative of con-

scious falsifying but are symptomatic of personality trends which might need probing; for the subject himself may be unaware of the motivational factors producing his false answers.

The "question score" is the total number of statements placed in the *Cannot Say* category. The authors of this inventory state that a high "question score" invalidates the others; for it is held that such high scores tend to move *high deviate* scores toward the mean. In other words, it appears that persons who tend to deviate from the mean (the normal range) more often classify statements as doubtful, whereas a correct classification would produce an even greater deviant score and rating. Here again the subject who has a high "question score" is not necessarily aware of the factors responsible for it.

The K score is a "correction factor" that is used to obtain increased validity of the scales. Application and interpretation of the inventory revealed that some normal persons got highly unfavorable scores in certain areas; scores that should indicate abnormality. An analysis was made of the items that were marked unfavorably by these normal persons (called "false positives"); these items constitute the K correction score. A low K score is said to indicate that the person was excessively severe in evaluating himself: over-candid, very self-critical, or exaggeration of minimal symptoms. A high K score, by contrast, represents a desire to make a more normal impression: unconscious or preconscious defensiveness against psychological weakness, or deliberate distortion. The authors of this inventory regard the K score as a measure of test-taking attitude; hence, it is a score that gives further evidence of the over-all validity of the scores on each of the several scales.⁸

The MMPI has been and still is one of the several most widely used inventories, both clinically and experimentally. As is to be expected when dealing with the subtle and often elusive problems of personality traits, clinical and experimental findings have not been in complete agreement; and some have been negative.

The authors of this inventory report in their manual that a high score on a scale correctly predicted the final clinical diagnosis in more than sixty percent of new psychiatric admissions. This is a very creditable performance, especially in view of the lack of agreement or uni-

⁸ See P. E. Meehl and S. R. Hathaway, "The K Factor as a Suppressor Variable in the Minnesota Multiphasic Personality Inventory," *Journal of Applied Psychology*, Vol. 30, 1946, pp. 525-561.

formity (low reliability) among psychiatrists themselves, who make the clinical diagnoses and classifications.⁹ As a result of these findings, it may be said, at least, that the MMPI is valuable in facilitating diagnosis and in describing and predicting expected behavior.

Other published studies report significant agreement between scale scores and hypochondriasis, paranoia, schizophrenia, and, in particular, depressions. Also, numerous investigations have reported that scale scores readily distinguish between pathological groups (undifferentiated), on the one hand, and normal persons, on the other. In other words, the Multiphasic inventory is more effective when its validity findings are not affected by the uncertainties and unreliabilities of current psychiatric classifications.¹⁰

The Cornell Index.¹¹ This inventory has been devised to provide an instrument “. . . for the rapid psychiatric and psychosomatic evaluation of large numbers of persons in a variety of situations.” The index “. . . was assembled as a series of questions referring to neuropsychiatric and psychosomatic symptoms, which would serve as a standardized psychiatric history and a guide to the interview, and which, in addition, would statistically differentiate persons with serious personal and psychosomatic disturbances from the rest of the population. It was devised as an adjunct to the interview, not as a substitute unless an interview is impractical.”¹² This questionnaire, standardized for males only, consists of one hundred and one items. The questions fall into two groups: those differentiating sharply between persons with serious personality disturbances (for example, “Does worrying continually get you down?”) and those concerned

⁹ The reader will recall that obtained validity of any test depends in part upon the reliability of the criterion.

¹⁰ See J. C. McKinley and S. R. Hathaway, “The Identification and measurement of the Psychoneuroses in Medical Practice: The MMPI,” *Journal, American Medical Association*, Vol. 122, 1943, pp. 161-167; H. G. Gough, “Diagnostic Patterns on the MMPI,” *Journal of Clinical Psychology*, Vol. 2, 1946, pp. 23-37; P. E. Meehl, “Profile Analysis of the MMPI in Differential Diagnosis,” *Journal of Applied Psychology*, Vol. 30, 1946, pp. 517-524; S. R. Hathaway and P. E. Meehl, *An Atlas for the Clinical Use of the MMPI*, University of Minnesota Press, 1951; P. L. Sullivan and G. S. Welsh, “A Technique for Objective Configural Analysis of MMPI Profiles,” *Journal of Consulting Psychology*, Vol. 16, 1952, pp. 383-388.

¹¹ By H. A. Weider, H. G. Wolff, K. Brodman, B. Mittelman, and D. Wechsler. The Psychological Corporation, 1944-1949.

¹² Authors' Manual.

with significant bodily symptoms (for example, "Do you usually have trouble in digesting food?"). The questions are undisguised and often extreme (e.g., "Are you keyed up and jittery every moment?" "Are you a sleep walker?"). They must be answered either *yes* or *no*.

The authors of the Index report that it has been effective in showing the presence of anxiety states, hypochondriasis, asocial trends, convulsive disorders, migraine, asthma, peptic ulcers, and borderline clinical syndromes. It is to be noted that this inventory, unlike the Bernreuter, the Minnesota, and others, does not provide separate scoring scales and norms for specific personality traits or disorders. Its scores for the entire inventory are intended only to assist in distinguishing between those having serious personality or psychosomatic difficulties and those not having them. The scoring of the inventory is to be followed by an interview, or interviews, after which the diagnosis may be made.

The 101 questions themselves have been classified under the following ten categories, the number of items varying from one to another: defects in adjustment expressed as feelings of fear and inadequacy; pathological mood reactions, especially depression; nervousness and anxiety; neurocirculatory psychosomatic symptoms; pathological startle reactions; other psychosomatic symptoms; hypochondriasis and asthenia; gastrointestinal psychosomatic symptoms; excessive sensitivity and suspiciousness; troublesome psychopathy. This is distinctly an instrument for clinical use, chiefly for screening purposes and expediting diagnosis.

Two unusual features characterize the scoring of this personality inventory: cut-off scores and "stop" scores, both based upon a total of 1000 cases at military installations. Of these, 400 were men who had been rejected for neuropsychiatric reasons, and 600 were men accepted after psychiatric interview. A table of cut-off scores shows: (1) the percentage of rejectees, at each score level, who would have been identified by the scale, and (2) the percentage of those accepted after psychiatric interview, but who would have been rejected by the Index (Table 48). Thus, a cut-off score of 13 would have identified 74 percent of those rejected after psychiatric interview, as well as having rejected 13 percent of the men passed after interview.

The "stop" questions (e.g., "Were you ever a patient in a mental hospital?") are such as would indicate extreme maladjustment or pathology. The "stop" items are to be used for ready identification of

TABLE 48

*Per Cent of Psychiatric Accepts and Psychiatric Rejects * Identified as Rejects at Various Cutoff Levels of the Cornell Index*

Cutoff Level	400 Psychiatric Rejects	600 Psychiatric Accepts
0	100%	100%
1	99	82
2	97	67
3	94	54
4	93	46
5	92	39
6	90	32
7	86	28
8	85	24
9	83	20
10	81	18
11	78	16
12	76	15
13	74	13
14	72	12
15	68	10
16	66	9
17	62	8
18	61	7
19	60	7
20	57	6
21	55	5
22	53	4
23	50	4
24	48	4
25	45	3
26	42	3
27	41	3
28	40	2
29	39	1
30	35	1
31	34	1
32	32	1

* In terms of opinion at psychiatric interview at five induction stations.

Table reads: If cutoff score of 7 on Index were used, 86% of those rejected at interview would have been rejected by Index; 28% of those accepted at interview would also have been rejected by Index. From Manual of Cornell Index. Psychological Corporation. (By permission.)

men who, presumably, are to be considered immediately for rejection.

The efficiency of this *Index* in identifying poor personality risks, as shown in Table 48, is great enough to warrant its use for the purposes stated by its authors, especially in situations where large numbers of persons must be rapidly screened. In situations where such pressure does not exist, the *Index* is still useful as a basis for and a guide to subsequent psychological interview and to psychotherapy. The fact that the *Index* does not identify larger percentages of probable poor risks at some levels, while, at the same time, it rejects a number of "psychiatric accepts," may be due to one or both of two factors: (1) some inadequacy of the inventory, and (2) errors in psychiatric judgments leading to acceptance or rejection after psychiatric interview. It is widely recognized that the brief psychiatric screening interviews during World War II were not optimally conducted; and the psychiatric interviewers, in many instances, were very inadequately prepared for their tasks.

The Guilford-Zimmerman Temperament Survey.¹³ This is one of the most carefully developed inventories available. It presents rather full statistical data and psychological rationale, thus enabling the prospective user to evaluate the instrument for his needs and to interpret the results more reliably.

Ten traits are measured: general activity, restraint, ascendance, sociability, emotional stability, objectivity, friendliness, thoughtfulness, personal relations, and masculinity. The inventory is intended for use with individuals in grades 9 through 16 and with adults. The particular traits included are the products of factorial analyses made over a period of years by Guilford and his associates.

The statements of these authors regarding the values and characteristics of their inventory are much more restrained and psychologically cautious than those of some other authors of inventories.

The SRA Youth Inventory.¹⁴ Differing from other inventories, this one contains a check-list of 298 items, subdivided into nine classes:

¹³ By J. P. Guilford and W. S. Zimmerman. Sheridan Supply Co., Beverly Hills, California, 1949.

¹⁴ By H. H. Remmers, B. Shimberg, and A. J. Drucker. Science Research Associates, 1950.

"my school, looking ahead, about myself, getting along with others, my home and family, boy meets girl, health, things in general, basic difficulty." It is obvious from this list that the inventory is not devised to measure personality traits (except as they might be inferred from the scores). The check-list of items is intended to identify areas of problems that are worrying junior and senior high school pupils most. The obtained results, then, in conjunction with other information about each individual, are to be used in counseling and interview, perhaps also in instances where therapy is indicated.

Original materials for the inventory were obtained by having high-school pupils write spontaneous anonymous essays on problems of greatest concern to themselves. The contents of these were analyzed; earlier investigations and inquiries regarding teen-age problems were surveyed; and on the basis of these analyses and surveys, a preliminary list of items was prepared. These items were edited by specialists in psychology and in education, thus deriving the final list of items from direct introspection by the persons concerned, from comparative studies, and from professional evaluations.

The manual of this inventory provides full statistical data essential to the interpretation of scores. Equally important is the authors' emphasis upon interpreting scores in conjunction with other relevant information about each individual. When all relevant information about a particular adolescent is available, it is then possible to estimate the significance of his responses and to differentiate between problems which are relatively superficial and those which suggest basic personality difficulties.¹⁵

The Security-Insecurity Inventory.¹⁶ This is another instrument that differs from most others in that it is devised to assess only one pair of opposed personality traits. The authors of this inventory have selected these particular traits because they believe that security "is almost synonymous with mental health." Security is defined, essentially, as feelings of being liked, loved, and accepted; of belonging and having a place in the group; of safety and of being unanxious. The S-I Inventory is intended for use with groups for research and

¹⁵ A similar type of device is the Mooney Problem Check List (1950), published by the Psychological Corporation. It does not yield any scores, being intended to facilitate and expedite psychological interview and counseling.

¹⁶ By A. H. Maslow et al. Stanford University Press, 1952.

TABLE 49

Information on Reliability and Validity of Certain Inventories

<i>Inventory</i>	<i>Reliability</i>	<i>Validity Criteria</i>
Allport A-S Reaction Study	.85-.90 (split-half) .78 (retest)	occupational groups ratings of judges
Bell Adjustment Inventory	.75-.97 (retest) .80-.89 (odd-even)	other inventories ratings of judges item analysis differentiation of extreme groups
Bernreuter Personality Inventory	.78-.92 (split-half)	other inventories differentiation of extreme groups low intercorrelations of part-scores
California Test of Personality	.51-.97 (for part-scores) .80-.96 (for total scores)	later clinical findings experts' judgments of item appropriateness
Aspects of Personality	.52-.92 (retest)	item analysis item analysis behavior ratings
Minnesota Personality Scale	.90 (odd-even)	item analysis extreme groups known groups of ad- justed and mal- adjusted students
MMPI	.56-.90 (retest; normal subjects) .52-.89 (retest; psychiatric patients)	diagnosed psychiatric groups normal subjects amount of score over- lap between nosolo- gical groups differentiation be- tween unselected patients and normal persons
Cornell Index	.95 (Kuder- Richardson for- mula)	neuropsychiatric cases: after interview normal persons, ac- cepted after inter- view civilian groups other inventories distribution of scores in a college popu- lation

TABLE 49 (Continued)

Information on Reliability and Validity of Certain Inventories

<i>Inventory</i>	<i>Reliability</i>	<i>Validity Criteria</i>
Guilford-Zimmerman Temperament Survey	.75-.85	low correlations between traits factor analysis to isolate traits
SRA Youth Inventory	.75-.85	spontaneous productions of subjects (teen-agers) biserial correlations intercorrelations of part-scores professional judgments of diagnostic value of each item scores of known groups other inventories self-estimate of subjects known groups systematic analysis of syndromes, security and insecurity
Security-Insecurity Inventory	.84 (retest) .86 (odd-even)	

survey purposes, and for the screening of college students who may be in need of psychological counseling or of therapy.

Summary Data. The foregoing instruments were described as representatives of many in the same category. They were presented, as well, for the purpose of providing the reader with a conception of types of items included and of the personality traits being evaluated.

Of the personality inventories developed specifically for use in the armed forces during World War II, no specimens or descriptions have been given here because in essentials of conception, construction, and content they are not different from those in civilian use before the war. In fact, many previously published tests were employed in the armed forces.¹⁷ Results reported on the effectiveness of these inventories in the armed forces indicate somewhat greater validity and usefulness in some instances than do similar data for inventories used in civilian

¹⁷ See J. P. Guilford, editor, *Printed Classification Tests*, Report No. 5 (U. S. Government Printing Office, Washington, 1947), Chapters 22-27.

situations. The principal reasons for this seem to be greater motivation, screened populations, knowledge of both test results and of the subjects on the part of the judges, and specially devised inventories for specific situations.

Table 49 presents in outline form information relevant to the reliability and validity of each of the inventories described in this chapter.

BIOGRAPHICAL DATA QUESTIONNAIRES

With this type of inventory, the purpose is to sample the several areas of a person's experiences that appear to be associated, directly or indirectly, with behavior and success in a particular occupation or situation. These aspects may include the intellectual as well as the nonintellectual: for example, education, religious activities, occupations, interest in athletics, marital record, financial status and interests, specific skills, group affiliations, family relationships, etc. The development of a biographical questionnaire involves a job analysis and experimental determination of biographical areas that are most significant.

Items of a questionnaire are intended to reveal the subject's attitudes and adjustments in terms of significant situations in which he has participated. This approach is not an attempt to measure personality traits. Its major premise is that past behavior, interests, activities, habits, skills, attitudes, etc., are indicators of what may be expected in the future. Actually, biographical data questionnaires are variations on the familiar application blank, including a wide range of information, with answers to be given in multiple-choice form.

During World War II, this technique was employed in the armed forces in an effort to improve the selection of personnel for various training assignments.¹⁸ The results were not particularly successful. They were, however, such as to indicate that the method has promise, although much further experimental work remains to be done.

The following is an item from the Guilford volume (footnote 17, p. 772).

Mark any of the following types of work which you have done at any time and for which you have received remuneration. (More than one may be marked.)

¹⁸ See D. B. Stuit, *op. cit.*, *passim.*; and J. P. Guilford (ed.), *Printed Classification Tests*, Chapter 27.

- A. Manufacturing industries (machine operator, factory hand, textile worker, etc.)
- B. Technical trades (baker, electrician, radio repairman, etc.)
- C. Transportation and communication (truck driver, linesman, deckhand, etc.)
- D. Business trades (store clerk, salesman, agent, window dresser, etc.)
- E. Public service (fireman, policeman, forest ranger, soldier, etc.)

A recent attempt to devise a biographical questionnaire for selection of educational administrators sampled information in the following areas: childhood and early background; professional preparation; health; interests; early signs of leadership; heterogeneous items.¹⁹ Two sample items follow:

During most of the time before you were 16 you lived:

- A. with both parents
- B. with one parent
- C. with a relative
- D. with foster-parents or non-relatives
- E. in a home or institution

How often were you a leader of your childhood "gang" activities up to the age of 12 years?

- A. always
- B. frequently
- C. occasionally
- D. seldom or never
- E. never a member of a gang; or can't remember.

While this investigation proved of little value for the selection of school administrators, it is reported here as representative of the approach to the problem and of the early stages of a technique which may prove fruitful for purely practical purposes.

The biographical data method has within it a very serious potential defect not found in other types of inventories. It is a purely empirical procedure, in which an attempt is made to find the predictive value of each biographical datum with respect to "success" in a specific position or in a type of occupation. Since this type of questionnaire does not deal with abilities, skills, and basic personality traits, and

¹⁹ J. P. Guilford and A. L. Comrey, "Prediction of Proficiency of Administrative Personnel from Personal-History Data," *Educational and Psychological Measurement*, Vol. 8, 1948, pp. 281-296.

since it is validated against job performance ratings and against "progress" of individuals in an organization, the predictive significance of biographical data is very largely a matter of the standards, values, and biases of raters or of the particular organization. For example, it may be found that being "rural-born" (rather than "city-born") has a fairly heavy positive weight in predicting progress to a school superintendency, or to an executive post in industry. This situation would indicate nothing about the subject's abilities for the position nor of his personality traits, nor of the actual demands made by the position itself. Such a finding would, however, assist in an analysis of the values and biases of prospective employers.

TESTS OF ATTITUDES AND VALUES

Characteristics. An attitude is a dispositional readiness to respond to certain situations, persons, objects, or ideas in a consistent manner which has been learned and has become one's typical mode of response. An attitude has a well-defined object of reference. For example, one's views regarding a class of food or drink (e.g., fish and liquors), sports, mathematics, or Democrats are attitudes. If, however, a person's characteristic behavior is described as self-sacrificing, intellectual, liberal—or the opposite of these—some of his *traits* are being indicated, since these terms represent his *generalized* ways of behaving and viewing situations. The degree or strength of a person's attitude may vary from extremely positive through a gradation to extremely negative. Obviously, it is possible to construct tests of innumerable attitudes.

Basically, tests of attitudes follow the principles suggested and employed by Thurstone for developing such scales.²⁰ Several assumptions are made: (1) the scale should deal with a controversial question; (2) an individual's feelings (and, it should be added, insights) in regard to the question will determine his responses to the various statements that are made pro and con; (3) the statements may be scaled regarding the degree to which they favor or are opposed to the question under consideration.

Methods of Scaling. Thurstone's technique of scaling attitude tests known as the method of *equal-appearing intervals*. The method is

²⁰ L. L. Thurstone, *Scales for the Measurement of Social Attitude*, Chicago: University of Chicago Press, 1930; "The Measurement of Social Attitudes," *Journal of Abnormal and Social Psychology*, Vol. 26, 1931, pp. 249-269.

essentially this. Statements, both favorable and unfavorable, bearing on a particular problem, question, or institution are obtained from a group of selected writers, other experts, and laymen. These statements are edited. Then they are classified by a large number of judges on an eleven-point scale. This is done by placing each statement in one of eleven piles, presumably forming a continuum, according to degree of favorableness or unfavorableness of each item with respect to the question at hand. The median of the judged locations for a statement (item) is its scale value. Statements which are judged to be ambiguous or irrelevant to the continuum are eliminated.

Before inclusion in the final scale, each question is analyzed for consistency with the general attitudes found by the total scale. For example, on a scale to determine attitudes toward churches, if it is found that many persons having an unfavorable attitude check a statement that is apparently favorable, then that item is considered irrelevant and is discarded. Statements having approximately the same values in the scale should show high consistency in degree of endorsement by each subject. This is essentially a simple method of item analysis. Ambiguity of an item is determined by the spread or range of judges' ratings in the original eleven-fold scale, given in terms of *Q* (quartile deviation). If an item's *Q* is "high," it is eliminated.

In taking an attitude test scaled in this manner, the respondent checks those statements with which he agrees, his score being the median of the scale values of the items he has marked. Thurstone held that scales constructed for different attitudes by this method permit direct comparison of the scores of any attitudes so measured. The validity of such comparison, however, has been questioned because the defined "neutral points" of different attitudes are not necessarily the same. Nor are the intervals demonstrably equal; they are only equal *appearing*. The Thurstone method is useful if strict comparability of scores is not assumed.

Thurstone and his students developed a series of scales, each consisting of statements from extremely favorable to extremely unfavorable. The topics included in these scales dealt with attitudes—among others—toward Negroes, Chinese, war, censorship, the Bible, patriotism, and freedom of speech. The following statements are from the scale of attitudes toward the church.²¹ The scale value of each

²¹ L. L. Thurstone and E. J. Chave, *The Measurement of Attitude*, Chicago: University of Chicago Press, 1929.

statement is given in parentheses, low values being favorable and high values unfavorable, with a possible range from 0 to 11.

I find the services of the church both restful and inspiring. (2.3)

I think the church is a parasite on society. (11)

I believe what the church teaches but with mental reservations. (4.5)

I think the teaching of the church is altogether too superficial to have much social significance. (8.3)

I believe in religion but I seldom go to church. (5.4)

I believe the church is the greatest institution in America today. (1.7)

Likert suggested the use of an attitude-scoring technique which is simpler than the Thurstone method and is regarded by many as being at least as reliable. Each item, or statement, in the attitude scale is followed by five responses, one of which is checked by the subject. The responses, indicating degree of strength of attitude, are: strongly agree, agree, undecided, disagree, or strongly disagree. (Approve-disapprove may be used in place of agree-disagree). Arbitrary scoring weights of 1, 2, 3, 4, 5 were assigned for the respective responses. An individual's score on a particular attitude scale is the sum of his ratings on all items. The principal advantage of Likert's method, obviously, is that it makes unnecessary the use of a group of judges to arrange statements into categories representing degrees of favorableness or unfavorableness. However, since the items are selected on an *a priori* basis, and since the scoring weights are arbitrarily assigned, the use of the Likert method, like the Thurstone, measures attitudes only in the sense that individuals are given a rank order according to attitude intensity.²²

The following items from the Minnesota Personality Scale (for men) are examples of the technique suggested by Likert.

SA., A., U., D., SD. On the whole lawyers are honest.

" The future looks very black.

" Education only makes a person discontented.

Remmers and his collaborators have prepared a series of attitude scales which differ in construction from Thurstone's in that each scale is intended to measure an attitude toward a larger group of objects,

²² See R. Likert, *A Technique for the Measurement of Attitudes*, Archives of Psychology, Vol. 22, No. 140, 1932.

persons, or institutions. The scales deal, among others, with national and racial groups, vocations, teachers, social action, and school subjects.²³

Tests of Values. A test of values, by contrast with one of attitudes, purports to measure generalized and dominant interests. The *Study of Values*,²⁴ for example, is based upon six types of values, as classified by Spranger. Their items are intended to measure the relative prominence of the subject's interests, for the purpose of classifying his values. The six types are: theoretical, economic, aesthetic, social, political, and religious. According to this classification, the dominant interest of the theoretical man is discovery of truth; the economic man is interested in what is useful; the aesthetic man values most form and harmony; the highest value of the social type is love of people; the political man is interested primarily in power; and the religious man places highest the value of unity in an effort to comprehend the cosmos as a whole. This test of values presents forty-five problem situations, under each of which the subject is required to select—from paired alternatives or from multiple choices—responses which are indicative of degrees of the six types of interests and values. For example:

The main object of scientific research should be the discovery of truth rather than its practical applications. (a) Yes; (b) No.

Do you think that a good government should aim chiefly at—[The following statements are to be ranked in order of preference.]

- (a) more aid for the poor, sick, and old
- (b) the development of manufacturing and trade
- (c) introducing highest ethical principles into its policies and diplomacy
- (d) establishing a position of prestige and respect among nations

It is not to be assumed that these six are "natural" types, or that they include all possible value groups, or that individuals can be classified entirely under one or another. As a matter of fact, most persons are a mixture of two or more of these value groups, some values being stronger and more dominant than others in each person. The six classifications were employed as starting points for the in-

²³ H. H. Remmers, *Attitude Scales*, Division of Educational Reference, Lafayette, Indiana: Purdue University, 1934-1936.

²⁴ By G. W. Allport, P. E. Vernon, and G. Lindzey. Boston: Houghton Mifflin, 1931-1951.

vestigation of complex views of life which, among others, serve to give unity and purposefulness to the mature person.

An interesting and different approach to the study of values of high-school and college students and adults is found in the *Sims SCI Occupational Rating Scale*.²⁵ This scale is devised "to reveal the level in our social structure—i.e., the social class—with which a person unconsciously identifies himself." The scale lists forty-two occupations, representing all levels of socio-economic status. The subject indicates whether persons following each of the occupations generally belong to the same, or to a higher, or to a lower social class than he does himself. Sims states that ". . . by examining the occupations which the subject indicates are those whose followers belong to his own social class, we are able to determine the position which he assigns himself in our society." This inventory may be regarded as one that estimates one set of social values; for affiliation with a socio-economic group usually signifies acceptance of the major values of that group.

Evaluation of tests of attitudes and values. There is little to choose between the methods of attitude testing, so far as reliability is concerned. On the various scales devised, the median reliability coefficient has been about .70; with some below .60 and even below .50, while others were above .80. In a few instances, the coefficients were so low as to be negligible. The two methods of scoring (Thurstone's and Likert's) are themselves very highly correlated (about .90), as shown when the same sets of statements and responses were scored both ways.²⁶

Validity of tests of attitudes and values is extremely difficult to determine in a statistical manner, since the only behavioral criterion is people's actions. Obviously, it is practically impossible to obtain objective behavioral data on a population sampling with regard to such matters as attitudes toward church, foreign-born, specific minority groups, education, and the like.²⁷ Furthermore, overt behavior need

²⁵ By V. M. Sims. World Book Co., 1952. SCI stands for "social class identification."

²⁶ A. L. Edwards and K. C. Kenney, "A Comparison of the Thurstone and Likert Techniques of Attitude Scale Construction," *Journal of Applied Psychology*, Vol. 30, 1946, pp. 72-83.

²⁷ Cf. S. M. Corey, "Professed Attitudes and Actual Behavior," *Journal of Educational Psychology*, Vol. 28, 1937, pp. 271-280; R. T. LaPiere, "Attitudes vs. Actions," *Social Forces*, Vol. 13, 1934, pp. 230-237.

not always be correlated with attitude scores. For example, consider two persons, one of whom is extremely hostile to churches while the other is indifferent to them. Their scores on the scale will differ significantly; but the indifferent person might attend and support a church as little as the hostile one. On the other hand, another indifferent person might attend regularly for social or economic reasons. Similarity of manifest behavior may be true, also, of one who is hostile to the foreign-born and one who merely avoids them. The fact that attitudes and overt behavior need not correspond makes validation, in the usual terms, a near impossibility. It is reasonable to conclude, however, that if individuals make a genuine effort to respond according to their own attitudes, these scales are useful in evaluating the beliefs of the respondents, as of the time the responses are given. Whether retest scores will show significant changes will depend upon each respondent's experiences, including educational, in the interim.

Scales of attitudes and values have had little use in clinics or in educational testing programs. They have been utilized principally in investigations in social and educational psychology.

OPINION POLLING

In recent years opinion gauging has become in itself a specialized field of study and practice. The most prominent activities have been in making political predictions, especially regarding presidential elections. But opinion polling has been concerned also with a great variety of subjects dealing with social, economic, international, military, and other questions, and also with questions of consumer preferences (usually called market research).

While some opinion studies use several questions in surveying a given issue, many employ only a single question. The question may be given in one of several forms: some require merely a "yes" or "no" answer; some require a rating of intensity or degree, such as "strongly approve," "approve," etc., or "very much," "much," etc.; at times the respondent is asked to check or rank items in a given list; sometimes the respondent classes one of two alternatives; occasionally the question is of the "open-end" type, in which the respondent completes a statement or sentence to suit himself.

All replies are classified into two or more categories, intended to answer the purpose for which the survey is planned. Usually the percentage of responses in each category is found. The responses in the

several categories are regarded by most opinion gaugers as representing quantitative or qualitative differences in attitude with respect to the question being investigated.

The mailed questionnaire, which had been in use long before opinion polling became popular, is another form of opinion gauging. This form of questioning, however, presents several serious disadvantages, so that it is not widely used at present. Representative mailing lists are difficult to obtain or develop; the percentage answering the questionnaires is often small and atypical of the total group; the questions may not be understood or correctly interpreted by persons at the lower end of the intelligence scale.

Opinions are obtained from a sampling of persons who are regarded as representative of a defined population. This population may be any group that it is desired to study: for example, college students, or farmers, or trade-union members, or housewives, or all the voters of the United States. Once the population to be polled has been defined, it may be sampled by one of three most common methods: (1) random sampling; (2) stratified sampling; (3) area sampling.

Random sampling simply means selecting respondents by "chance," with the expectation that each individual in the defined group has an equal chance of being chosen. Hence, it was assumed that random selection of a large enough number of individuals would yield a representative sampling of the entire group. It has been shown, however, that very often completely random selection does not yield a representative group. Precautions must be taken to avoid obtaining a biased sampling; and the investigation must be so designed as to obtain a random sampling within the design. In other words, completely unplanned, haphazard polling does not insure that a representative set of responses will be obtained. Consequently, purely random sampling of respondents is infrequently used at present.

When *stratified sampling* is used, the population to be studied is classified into two or more categories (or strata), from each of which a sample of proper proportions is drawn. For example, it might be desired to divide the population into two groups: high-school graduates and non-high school graduates. The proportions of each in the general population would have to be determined; then the same proportions would be sampled from each group in the investigation.²⁸ Finer subdivisions might be desired, such as completion of six grades,

²⁸ In some instances, for well-defined reasons, disproportional representation is warranted.

and nine, and twelve, and college graduation. In other situations, it might be necessary to obtain respondents who satisfy specified proportions with regard to sex, age, economic status, native or foreign-born, etc., or combinations of two or more of these. It is obvious that the problem of stratified sampling becomes more complex and difficult as the number of characteristics is increased. The purpose of the investigation determines the characteristics upon which the sampling shall be "stratified." The population to be studied is then analyzed for these characteristics and sample respondents are drawn accordingly.

Using the method of *area sampling*, the geographic area to be investigated is divided into a number of subdivisions of approximately equal population. The characteristics of the population in the various areas are known, so that in the final selection, individuals of the desired characteristics will be included in correct proportions. The particular divisions to be studied are chosen so as to yield a sampling that will be as large and diversified as may be necessary for the investigation. Each of the divisions thus included is further subdivided into much smaller units, within each of which are only a few persons who have the desired characteristics of those to be polled by interview. Then from among these small subdivisions, a random sampling is made for actual study. Within each of these final units—chosen by means of random numbers—every person is interviewed, even if repeated visits to the homes are necessary.

The area sampling method is infrequently used because it is so expensive; it requires that relevant information be obtained regarding every family within specified areas and that extremely elaborate files be kept. It is very doubtful whether the advantages of the area sampling method are such as to warrant the cost and the attempts to obtain information of kinds that many people will, with justification, regard as an invasion of their privacy.

Polling procedures and results are too often of doubtful reliability and validity because of difficulties inherent in the practice itself and because some of the problems have been neglected or inadequately dealt with. The major problems and difficulties will be indicated without elaboration.

With any method of sampling, it is extremely difficult to get a completely unbiased sampling because of unknown chance errors, unknown selective factors, and errors of judgment in evaluating traits and responses of some persons.

It is doubtful if opinions can be correctly gauged by a single question, except in special instances such as elections, when the respondent makes a choice between candidates.

Answers to questions may be deliberately falsified or there may be lack of frankness which results in a large number of "undecided" or neutral answers.

All respondents do not necessarily interpret a question or statement in the same way; the same question or statement has varying connotations for different persons.

Some respondents do not understand the *language* or phrasing of the question or statement.

Some respondents do not know or understand the *issue* being dealt with.

Individuals who do not actually have an opinion feel at times constrained to express one anyhow.

Over-all percentages of responses vary with different ways of stating a question.

When the open-end question is used, it is difficult to classify the responses.

Responses are influenced by the training, influence, and possible bias of the interviewer.

Significant differences in status between interviewer and respondents may influence answers: e.g., responses of Negroes to Negro may differ from those of Negroes to white.

Verbal expression of an opinion does not necessarily indicate the respondent's actions.

Different persons may have different reasons for giving the same responses; and persons giving different responses may do so for a common reason. Psychologically, it is necessary to determine, through skillful interview, the reasons for a response.

Some of these difficulties and problems can be met in part if the questions are stated unambiguously and simply and are easily understood by persons for whom intended; if the respondents are familiar with the issue; if the respondents could reply by "secret ballot" or could surely remain anonymous; if more than one question is used for a given issue.

As psychologists, however, we are primarily concerned to know *why* individuals hold certain opinions regarding such issues as are polled, rather than to learn only that certain percentages of the sampling think thus or so. Why do certain persons choose to vote as they do? Why do a large majority of a certain category of housewives prefer pastel colored refrigerators to white? Why do certain persons believe that a certain foreign country, X, is a threat to the United

States? Why are some individuals hostile to persons of Oriental origin? Determination of reasons for these and even more subtle behaviors, attitudes, and values requires interviewing by qualified psychologists or other qualified professional persons. So long as opinion polling remains a matter of classification of responses and determination of percentages, it remains a statistical problem of significance, aside from technical considerations, principally to political scientists, sociologists, and consumer research specialists.²⁹

EVALUATION OF PERSONALITY INVENTORIES

Reliability and Validity. The authors of most of these instruments present reliability coefficients which are reasonably satisfactory, or even quite high, being for the greater part .80 or more. In some instances the reliabilities falling below .80 or even below .70 are unsatisfactory so far as predictions of subsequent behavior in individual cases are concerned. The methods used are the usual ones: namely, correlation of odd-even scores and of test-retest scores.

It is in respect to validity, however, that personality inventories as a class present the greatest difficulties and are most vulnerable to criticism. Determination of validity is certainly difficult; yet that must be the most essential requisite of a useful instrument.

In devising their personality measures, the earlier authors began with questions or statements, gathered from a variety of publications and sources (clinics, schools, colleges, industry and business, home, community), that are symptomatic of neurotic disorders, behavior difficulties, or of normal behavior manifestations. The nature and scope of the items depend, of course, upon the age-levels and purposes for which a particular inventory is intended. Authors of later instruments followed the same practice, often utilizing items from several earlier devices, recombining them, and adding some new ones.

Methods. The following methods and criteria have commonly been used in studying validity: (1) statistically significant differences between average scores of clinically well-defined groups; (2) significant average score differences between clinical groups and a normal population; (3) ability of each item to differentiate between the two extreme groups in the standardization population; (4) internal con-

²⁹ For a comprehensive critique of opinion polling see: Q. McNemar, "Opinion-Attitude Methodology," *Psychological Bulletin*, Vol. 43, 1946, pp. 289-374.

sistency of items or parts; (5) comparisons of inventory scores with judgments of counselors and school officials; (6) selection of items from other published tests and correlations with these instruments; (7) factorial analysis; (8) the author's own judgment regarding manifestations which constitute evidence of a specified trait.

(1) The first criterion should be employed only when an inventory is designed primarily for clinical use in the diagnosis of personality disorders; as, for example, in the case of the Minnesota Multiphasic Personality Inventory (depression, hysteria, paranoia, etc.). Measures standardized on this criterion cannot and should not be used for the study of a normal population, except for the purpose of screening out, for further clinical study, individuals at the extremes of maladjustment.

(2) The second criterion is likewise used with measures that are intended chiefly for clinical purposes, but in this instance the emphasis is upon segregation of the normal from the abnormal, rather than differentiation among the abnormal themselves.

(3) The third criterion evaluates each item in regard to its effectiveness in distinguishing between the extreme groups of a distribution of scores for a single trait (e.g., radical-reactionary or ascendance-submission), as shown by the percentage at each extreme answering the item in a specified manner. A test so validated should not be used for a general, representative population because it is not necessarily adequate to differentiate among the great percentage of persons who are located between the extremes.

(4) The fourth, internal consistency, differs from the preceding criterion in that each item is correlated against part scores for all subjects, the purpose being to learn whether answers to the individual items are, on the whole, reasonably consistent with the behavior or personality trends suggested by the scores. This is actually a form of "face validity"; for basic to it is the assumption that the total or part score actually does measure what it purports to, and that it is the author's task to eliminate those particular items that do not conform to his selected traits and to the test items as a whole. With few exceptions, it is very doubtful if internal consistency may be regarded as a measure of validity unless external criteria are used in addition. If this is done, then internal consistency will be sought only in an effort to get total scores which yield the highest validity coefficients against external criteria.

(5) The fifth criterion, employed in constructing inventories to be used principally in schools, assumes that the obtained judgments have adequate validity and that the judges are competent to assess personality traits as well as desirable and undesirable forms of adjustment. In some instances the assumption is warranted; in many it is not.

(6) The sixth criterion assumes that items and tests already in use are themselves valid. This is a practice which in many instances is not justified and which tends to perpetuate inadequacies, errors, and misconceptions inherent in the older inventories.

(7) Factor analysis, the seventh criterion, has in the minds and work of some investigators come to take the place of validation against behavioral and psychological analysis. These investigators assemble a number of items, administer the inventory to a standardization population, statistically analyze the scores, group the items into a number of categories, and give the categories names of traits that appear to be measured by means of the items that they decided, more or less arbitrarily, should go into the inventory in the first place. This is a form of circular reasoning that does not add to the construction of valid instruments.³⁰ Actual behaviors of defined groups of persons must be the ultimate criteria of validity of practically all personality inventories.³¹ For personality traits derive their ultimate significance from the role they play in advancing or retarding people's personal and social adjustment.

(8) In using the eighth criterion, an author simply selects or devises

³⁰ A valuable criticism of factor analysis in the study of personality is to be found in G. W. Allport, *Personality*, pp. 242 ff. On the techniques of factor analysis, see J. P. Guilford, *Psychometric Methods*, New York: McGraw-Hill Book Co., 1936, Chap. 14. As an example of an extreme position on factorial analyzed personality traits, see R. B. Cattell, *Personality: a Systematic Theoretical and Factual Study*, New York: McGraw-Hill Book Co., 1950. Factorial analysis has not yet succeeded in achieving a significant degree of simplification or integration of personality traits. One publication brings together comparable analyses and emerges with forty-nine personality factors. See J. W. French, *The Description of Personality in Terms of Rotated Factors*, Princeton, N. J.: Educational Testing Service, 1953. Also, R. B. Cattell et al., "The Objective Measurement of Dynamic Traits," *Educational and Psychological Measurement*, Vol. 10, 1950, pp. 224-248.

³¹ The possible exceptions to this statement are the values and attitudes scales. As already explained, one's expressed values and attitudes, as measured on a scale, may be genuine; yet for other, overbalancing reasons his actions may not be consistent with them.

items to suit his own definition of a trait or a theory of personality, without concern thereafter over their behavioral or statistical validity. Starting with theories and definitions is, of course, desirable; but the validation process must go beyond that stage.

Research Findings. It is not to be assumed that every test of personality, or many of them, for that matter, have been validated according to all or even several of these criteria. Most have been subjected to validation by the method of internal consistency, or correlation with earlier tests, plus, in some instances, the use of known groups, in one form or another. Validation data obtained by correlations of internal consistency have yielded the most impressive results. But this is readily understandable, because items can be retained, modified, or eliminated so as to create the desired internal relationship without, however, giving any assurance that the specified traits are actually being measured. Results obtained by intercorrelations of inventories among themselves, though high or moderate in some instances, have not, on the whole, been very satisfactory. Poorest results have been found in studies of validation against competently determined group classifications (known groups). Yet it is this method which is the most significant and crucial one. Using this criterion, experimental studies have yielded contradictory results.³² For example, of nine investigations to validate personality questionnaires with groups of behavior-problem children, the number of correlation coefficients of various sizes were as follows:

two above .70
one between .40 and .70
six below .40

In 75 validating studies, correlating scores of normals and abnormals (diagnosed neurotics and psychotics) with selected criteria, the following coefficients were obtained:

thirty-six above .70
nine between .40 and .70
thirty below .40

³² See A. Ellis, "The Validity of Personality Questionnaires," *Psychological Bulletin*, Vol. 43, 1946, pp. 385-440. This is a comprehensive summary of published studies on the subject and a valuable criticism of the most common practices. See also, by Ellis, "Personality Questionnaires," *Review of Educational Research*, Vol. 17, No. 1, 1947, Chapter 4.

When inventory scores were validated against ratings by teachers, friends, or associates, the findings were:

twelve above .70
ten between .40 and .70
twenty-two below .40

Validation studies of four group inventories (Bell, Bernreuter, Thurstone Personality Schedule, Woodworth Personal Data Sheet) yielded the following results:

twenty-five above .70
eleven between .40 and .70
forty-four below .40

More consistent and convincing results were obtained when personality tests—principally the Minnesota Multiphasic—were administered individually rather than to groups. The validity coefficients were:

ten above .70
three between .40 and .70
two below .40

These last data suggest that individual testing of personality is superior because subjects may be more highly motivated, due to clinical rapport; the inventories are more carefully developed; their uses are more limited and more clearly defined.

Somewhat more than half of the coefficients reported above (.40 and higher) are either quite high or moderate as validating data. And somewhat fewer than half are quite low (below .40). Although coefficients below .40 or .50 do not have very much predictive value for all individuals within the group, they may, nevertheless, indicate that the inventory has considerable value in identifying individuals who constitute the more deviant groups.

The differences found among the large number of studies summarized cannot be attributed to the inventories alone. Other factors to be considered are: the number of subjects, their homogeneity, and their classification; the soundness of the ratings or of the clinical diagnoses that are used as validity criteria; the purposes for which and the conditions under which the inventories were administered.

These findings indicate that inventories for the assessment of personality traits should not be used indiscriminately or uncritically; nor should the sounder among them be rejected uncritically. Personality inventories are more valuable for certain defined populations than for others; they are more valuable in some kinds of situations than in others. One comprehensive survey of published studies between 1946 and 1951 reports ". . . that in most cases inventory scores discriminate significantly when used with psychoneurotic, psychosomatic, alcoholic, age, sex, ethnic, and college groups. . . . and they usually do not give significant group discriminations when used with vocational, academic, socio-economic, and disabled and ill groups."³³

Factors Adversely Affecting Validity Findings. Many reasons have been offered in explanation of the equivocal and less satisfactory validity data reported in the numerous published studies. The major reasons will be briefly noted here.³⁴

The questionnaires sample segments of the person; they do not bring out the whole patterned, or organismic, representation of behavior. Personality cannot be described in terms of separate traits or a mere summation of traits. None of the available questionnaires, rating scales, and personal-history records are able to portray the personality as a complete, dynamic, organized whole. They measure—not very precisely—certain aspects of behavior. They do not actually measure or assess the unit (the organism) that does the behaving.

Many studies of validity have been devoted to correlations with differential diagnoses as criteria; but the psychiatric descriptions and classifications are not always clearly defined or sufficiently distinct; psychiatric diagnoses are often not sufficiently reliable; and many clinical subjects are too unstable or unresponsive in the test situation.

In testing for traits common to a population, attention may be diverted from the individual as a unit to the assignment of a mere rank or index to a segment.

³³ A. Ellis, "Recent Research with Personality Inventories," *Journal of Consulting Psychology*, Vol. 17, 1953, pp. 45-49.

³⁴ Cf. A. Ellis, *Psychological Bulletin*, 1946, *loc. cit.* Also, R. M. W. Travers, "Rational Hypotheses in the Construction of Tests," *Educational and Psychological Measurement*, Vol. 11, 1951, pp. 128-137; J. C. Flanagan, "The Use of Comprehensive Rationales in Test Development," *ibid.*, pp. 151-155; C. E. Osgood and G. J. Suci, "A Measure of Relation Determined by Both Mean Difference and Profile Information," *Psychological Bulletin*, Vol. 49, 1952, pp. 251-262.

Some inventories purporting to measure two or more separate traits are measuring largely the same trait under more than one name.

Differences in cultural factors will cause subjects to respond differently to the same question.

A given question or statement does not have the same meaning for all subjects, even when clearly stated. It is a fallacy to assume that all persons have similar reasons for giving similar responses to an item.

Misunderstandings of questions are due to vocabulary limitations of some respondents.

Many questions cannot be answered in the *yes, no, ?* form.

There is a general tendency for some subjects to overrate themselves ("self-halo").

Almost anyone can falsify his replies to a questionnaire; and an indeterminate number do so.³⁵

Some subjects lack insight into their traits; others fundamentally and unconsciously may be different personalities from their own conscious self-appraisals.

The scoring of answers to items is often based upon the test author's own personal judgments and set of values.

On some questionnaires, either very low or very high scores, or both, may be significant; but the wide middle range of scores may not be at all meaningful for differentiation and description.

Statistical assumptions and procedures often take the place of behavior analysis and psychological insights.³⁶

Positive Factors. On the positive side, the following points may be made.

Personality testing is a relatively recent practice and is still in process of development.

Efforts to develop measures of personality traits encourage greater uniformity in and precision of trait definition and description.

When there is essential agreement in regard to definition of traits and terms, and in regard to behavior and symptoms, the use of standardized inventories increases the objectivity of personality ratings and descriptions.

The use of personality measures encourages analysis of traits into their constituent elements, thus providing a better understanding of

³⁵ V. H. Noll, "Simulation by College Students of a Prescribed Pattern on a Personality Scale," *Educational and Psychological Measurement*, Vol. 11, 1951, pp. 478-488.

³⁶ Personality tests of several kinds were used extensively in the armed forces during World War II, seemingly with greater success than in civilian life. Both spurious and legitimate factors account for this. See the searching analysis by A. Ellis and H. S. Conrad, "The Validity of Personality Inventories in Military Practice," *Psychological Bulletin*, Vol. 45, 1948, pp. 385-426.

each trait. (The elements themselves, taken separately and in isolation, are not, however, the trait.)

In some cases when, consciously or unconsciously, persons misrepresent themselves by their answers on an inventory, the instrument may still be clinically valuable, because the fact that they have misrepresented is significant in understanding their personalities, by means of subsequent interviews.

Psychometric analysis is useful as one of several clinical procedures, when its results are considered in conjunction with other evidence (e.g., the individual's history and psychological interview).

Answers to items of a questionnaire may be employed as the starting points of subsequent psychological interviews, since answers to various questions and responses to various statements may be significant in themselves, or they may reveal significant patterns of behavior, attitudes, and feelings. In such instances the numerical scores and percentile ranks can be disregarded. Pragmatically, at the present time, a useful test of personality is one whose score or responses to individual items assist in identifying areas of actual or potential maladjustment for purposes of further, more intensive study and subsequent treatment. Conversely, they can help in the identification of areas of wholesome adjustment. At their present stage of development, this is one of the most useful ways in which results of personality inventories can be employed.

Personality inventories are useful in the study of group trends; that is, in differentiating between groups of adjusted and maladjusted rather than between individuals.

Since personality traits and attitudes may undergo change, it is to be expected that inventories will be less reliable (in terms of test-retest scores) than, for example, tests of general intelligence.

Concluding Statement. Testing, as a method in the study of personality, is a justifiable procedure. But the instruments at present available still permit too wide a margin of error for widespread or uncritical group use; nor are their results refined enough to warrant the elaborate statistical processes to which some of them have been subjected. Tests of personality should be employed only by professional personnel who are versed in the principles of their construction, who know their limitations, and who are capable of insightful psychological analysis of behavior.

Much more research is needed in this field. In addition to inventory reliability, the variables of personality tests will have to be of greater psychological significance. That is, the traits being measured should be more carefully defined, and every item should be clearly and significantly related to the chosen trait. The wording of items should

be clearer so that their meanings, so far as possible, may be uniform for all subjects. Improved, more finely graded means of answering should be provided in place of the categorical *yes*, *no*, *?*. The diagnostic value of each item should be established. For this purpose, statistical analysis is an essential tool.

The criteria against which personality tests are to be validated must also be made more reliable than they are at present. If, for example, clinical diagnoses are used as a criterion, they must be accurate and dependable. Too often this is not the case. In some instances "tentative diagnoses" instead of final ones have been employed, or, as not infrequently happens, diagnosticians do not agree among themselves.³⁷ Again—as another illustration—it is unsound to use a blanket classification such as "delinquency" or "problem behavior" as a criterion, because there are a variety of kinds of delinquents and problem behaviors, differently motivated and occurring under varying conditions.

In this chapter, the reader has become acquainted with the range of traits being tested, the general purposes of personality inventories, their similarities and differences, and the types of items being used. But in view of inadequacies of available questionnaires and of their inherent deficiencies, psychologists have been giving increasing attention, in recent years, to the study of personality by means of projective methods, which are presented in the following chapter.

³⁷ For example, P. Ash, "The Reliability of Psychiatric Diagnoses," *Journal of Abnormal and Social Psychology*, Vol. 44, 1949, pp. 272-276.

PROJECTIVE METHODS: THE RORSCHACH AND THE THEMATIC APPERCEPTION TESTS

DEFINITION AND EXPLANATION

Psychologically, the term *projection* means: (1) the unconscious process whereby an individual attributes certain thoughts, attitudes, wishes, emotions, or characteristics to objects in his environment or to other persons. (2) Projection also takes the form of attributing one's own needs to others in his environment. (3) Or it may take the form of drawing incorrect inferences from an experience. The process is not recognized as being of personal origin, with the result that the content of the process is experienced as an outer perception and of external origin.

A projective test, then, is one that provides the subject with a stimulus situation giving him an opportunity to impose upon it his own private needs and his particular perceptions and interpretations. The several forms of the projective method (pictures, ink-blot, incomplete sentences, word associations, one's own writings and drawings, and others) are intended to elicit responses that will reveal the individual's "personality structure," feelings, values, motives, characteristic modes of adjustment, or "complexes." He is said to project the inner aspects of his personality through his interpretations and creations, thereby involuntarily revealing traits that are below the surface and incapable of exposure by means of the questionnaire type of personality test.

Personality inventories are standardized questionnaires which ask how the subject feels or acts in a variety of representative situations. A projective test, by contrast, is unstructured; instructions are general and are kept at a minimum to permit variety and flexibility of responses; there are no suggested responses; the responses are the subject's own spontaneous interpretations or creations. (A partial exception to this is the "projective questionnaire," which is partially structured.) Responses to the projective test involve cognitive factors—those that relate to what is there to the senses—and affective factors, or feelings about what is there.

The purpose of the projective test is not told to the subject, a frequent practice being to inform him that it is a test of his imagination. The most widely used projective methods (e.g., the Rorschach inkblots, the Murray pictures, and the like) are tests of perception,¹ which is an individual process. The less clear-cut the situation, the greater will be the individual differences in perceiving it. These tests, therefore, provide relatively unrestricted opportunity for the exercise and expression of individual differences in perception; for each subject sees what he himself is disposed to see and does what he is personally disposed to do. In so doing, through his manner of confronting and responding to the stimulus situation, the individual indirectly tells the examiner about himself.²

By contrast with inventories that attempt to portray the personality segmentally and imply that personality is the sum of the segments, a projective technique attempts to view and understand a personality as a whole, and to interpret its components in their interrelationships. This viewpoint is generally known as global or organismic theory, according to which the whole and its parts are mutually related, the whole being as essential to an understanding of the parts as the parts are to an understanding of the whole. According to the organismic view, furthermore, the measurement and evaluation merely of components does violence to the whole organized structure.

¹ Normal perception is defined as awareness of objects, conditions, and relationships as unified, articulated mental structures. Perception is also defined as a mental complex or integration that has sensory experiences as its core. Disturbances of perception will be shown by lack of integration, distortion, and bizarreness.

² This is not to say that each person's perceptions and responses are completely idiosyncratic. To each stimulus situation there are certain responses that are quite frequently obtained from certain groups. But there are also numerous individual variations and combinations which give each person's total response-pattern a degree of uniqueness.

The organismic conception of personality has emerged from clinical studies of numerous individuals whose behavior could be understood only in terms of the interrelationships and interdependence of traits, and from the many experimental investigations of perception and behavior. Projective methods, therefore, are regarded by many psychologists as the most valuable tests of personality because they are concerned with complex mental processes and portray the whole personality.

THE RORSCHACH TEST

Description and Procedure. This is the well-known and widely used inkblot test, named after Hermann Rorschach, a Swiss, who began his experimentation with inkblots as a means of stimulating and testing imagination. He was not the first investigator to perceive the possibilities of inkblots in experimental psychology, although his work was the most extensive of any, continuing from 1911 to 1921. He is credited with being the first to develop a technique for their use in personality diagnosis.³ He also changed the emphasis from content analysis to determinant analysis, which is explained in the following pages. Rorschach developed his test and methods as a practical tool to be applied to clinical cases in the study of unconscious factors in perception, and to reveal dynamic factors of behavior and personality. He proceeded on the principle that every performance of a person is an expression of his total personality, the more so if the performance is concerned with nonconventional stimulus-situations in response to which one cannot wilfully conceal his individuality. In responding to inkblots, the subject is generally unaware of what he reveals by the reports of what he sees. Yet in telling what he perceives, the individual is giving a general portrayal of his personality.

The Rorschach Test—now used from the nursery-school level through all ages of adulthood—consists of ten cards, on each of which is one bisymmetrical inkblot. Five are in black and white with differently shaded areas. Two contain black, white, and color in varying amounts; three are in various colors (“chromatics”).

The cards, each of which is numbered, are presented to the subject

³ For a bibliography of early investigations, see J. E. Bell, *Projective Techniques*, New York: Longmans, Green, 1948, pp. 75–76. The great bulk of research on the inkblot technique has been published since 1930, more especially since 1935.

one at a time and in prescribed sequence. The instructions are very simple; the subject is asked, according to Rorschach's own formula: "What does it look like? What could this be?" Several clinicians and investigators, who have used the test extensively, have somewhat modified the original instructions, though not in its essentials. Klopfer and Kelley, for example, use this formula: "People see all sorts of things in these inkblots; now tell me what you see, what it might be for you, what it makes you think of."⁴ The principal differences in the directions of the several specialists who have evolved their own formulas is in the amount of encouragement and urging used to elicit from the subject the fullest possible response to each card. These modifications—in contrast with the rigid directions to be followed in administering tests of intelligence, specific aptitudes, etc.—might account in part for some of the inconsistencies found in the numerous published reports on the Rorschach.

Rorschach did not impose time limits; nor do most of the present users. Nor is there any fixed number of responses for each card. The examiner makes note of various aspects of the subject's behavior: namely, a verbatim record (so far as possible) of the responses; time elapsed between presentation of each card and the first response to it (reaction time); length of time in long pauses between responses; total time required for each card (response time); position in which card is held for each response (indicating extent of the subject's exploration of the stimulus-situation); the subject's extraneous movements and other behavior of significance. The three recordings of time are useful in determining emotional blocking or resistance to

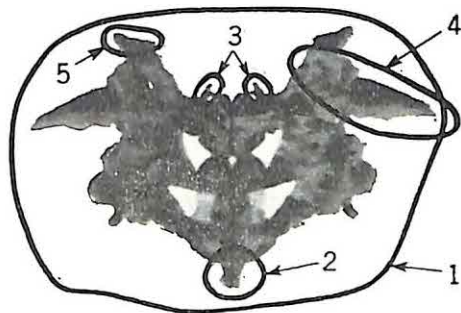


FIG. 19.1. A Rorschach Inkblot with Location of Responses. 1. Butterfly. 2. Pair of legs. 3. Hands pointing. 4. Profile of a dog. 5. A pointed cap. Inkblot without location, from H. Rorschach's *Psychodiagnostic plates*. (By permission Grune and Stratton.)

⁴ B. Klopfer and D. M. Kelley, *The Rorschach Technique*, New York: World Book Co., 1942. See also, S. J. Beck, *Rorschach's Test*. Vol. 1, *Basic Processes*, New York: Grune and Stratton, 1944.

what the individual might be perceiving in a particular inkblot.

Directly after all ten cards have been presented for responses, a second phase, the *inquiry*, follows.⁵ There are two main purposes of the inquiry. The first purpose is to learn just what parts or aspects of the blot were perceived and responded to; that is, what features of the stimulus-situation initiated and sustained the spontaneous association process: response to wholes, parts, small details, location, color, shading, apparent movement—all of which are essential items of information for scoring purposes. Second, in some instances the inquiry gives the subject an opportunity to add to or clarify his original responses; but if this is done, it must be completely spontaneous on the part of the subject and without any suggestion whatever from the examiner. The only questions that should be asked are those needed to clarify the scoring and to include in the score all the significant aspects of the responses. Too many or too leading questions by the examiner may elicit from the subject responses which do not represent his own perceptions but spring rather from suggestions more or less implicit in the questions asked. The only specific questions asked are for the purpose of determining the exact area or location to which particular responses refer, this being necessary for scoring and analysis under the several categories of the test.

Scoring. The scoring, following Rorschach for the most part, is based upon four major categories.

Location. The first is the *location*, or area, which has been perceived as the basis of each response. This may be the entire inkblot, a large portion, a small portion, a minute detail, or part of the white background. The area may be well defined, or merely vague and blurred. Location of responses is the basis of obtaining scores for wholes (called W's), large usual details (called D's), and small usual details (called d's), unusual detail (called Dd), and the white spaces (called S), which are parts of each person's pattern of response to

⁵ Some examiners prefer to conduct the inquiry immediately after the presentation of each card; e.g., M. R. Hertz, "The Method of Administration of the Rorschach Inkblot Test," *Child Development*, Vol. 7, 1936, pp. 237-254; D. Rapaport, et al., *Diagnostic Psychological Testing*, Chicago: The Yearbook Publishers, 1946, Vol. 2, pp. 97 ff. Various expert users of the Rorschach have introduced an occasional innovation of their own in administering and scoring the test, but these innovations will not be described. Only the basic aspects will be presented, without going into the "variations on a theme," so to speak.

the entire test. There are also other symbols used to designate other aspects of location; but these five are the major categories.

The locations of responses and the subject's ability to delineate them are regarded as indicative of his perceptual organizing processes, of his ability to analyze and articulate the parts, and of his associations as his perceptions shift within each blot. Analysis of responses in respect to location is said to reveal extent of the subject's perceptual organization or disorganization measured in terms of agreement with norms of perception and ability to analyze the whole and synthesize the parts.

Determinants. The second category includes the *determinants*, or characteristics, of the inkblot as perceived by the subject. The determinants are those aspects or qualities of the blot that have produced the responses to it. These may be *form*, *shading*, *color*, *perspective*, or *motion*—or combinations of these. Forms may be perceived with ordinary accuracy (F); or they may be unusual and clear percepts (F+); or poor percepts (F-). Generally, evaluation of form is a matter of the examiner's judgment, although some investigators have provided normative descriptions and numerical scores.⁶

The frequency, intensity, and interpretation of shading noted by the subject are recorded. The manner in which the subject responds to the shading (K) of the blots is said to be relevant to the manner in which he meets and satisfies his own affectional needs: whether by conscious denial of affectional need, or by a repressive mechanism, or by insensitivity and undeveloped affectional relationships with other persons.

With regard to color (C), the examiner records the particular colors reported and the manner in which the subject combines color with form, there being three categories: responses to pure color without form being involved (recorded as C); responses to a combination of form and color, in which form is dominant (FC); and responses to a combination of color and form in which color is dominant (CF).

A score for movement is given by most examiners when the subject

⁶ See B. Klopfer and H. H. Davidson, "Form Level Rating: A Preliminary Proposal for Appraising Mode and Level of Thinking as Expressed in Rorschach Records," *Rorschach Research Exchange*, Vol. 8, 1944, pp. 164-177. Also, Rapaport, *op. cit.*, pp. 107 f.; B. Klopfer et al., *Developments in the Rorschach Technique*, New York: World Book Co., 1954, Chap. 8.

perceives something going on in the blot, whereas Rorschach himself restricted the movement score (M) to responses which indicated empathy; that is, a true experiencing of or identification with the movement reported (obviously an extremely difficult phenomenon for the examiner to discern). At present a common practice is to reserve the symbol M for human movement, to designate animal movement as FM and inanimate movement as *m*.

The subject's mention or use of perspective or depth (FK) is also noted and scored. Parts of the inkblot are perceived as having perspective and being three-dimensional. FK, "in reasonable numbers," is said to be related to good adjustment, through attempts to handle affectional anxieties by introspective efforts.

The foregoing are the principal determinants as developed by Rorschach, and as further developed and modified by Klopfer and his collaborators. Somewhat different sets of determinants have been developed from Rorschach's original ones by several other psychologists, notably S. J. Beck.⁷ Although there are some differences in symbols used and in details of response classification, the conceptual similarities between the several systems are much greater than the dissimilarities. One of Beck's categories, in particular, should be mentioned in connection with response determinants; that is what he calls *organization* (designated by Z). This determinant, an extension of the concept of the whole (W), indicates ability to perceive (or create) "new and meaningful relations between portions of the figure not usually so organized." Organization is said to be related to level of intelligence. A relatively low Z rating, it has been observed, may be useful in detecting depression, anxiety, or unresponsiveness in persons known to be of superior intelligence.

Content. The third scoring category is *content*. Here the subject's responses are classified into the several more common groups such as plants, animals, humans, landscapes, man-made objects, anatomy, sex, and others. Content items are not merely classified into specified groups; they are used by the examiner as a source of ascertaining the subject's personal meanings, attitudes, interests, and even "complexes." Some examiners have interpreted content items, also, as having psychiatric or psychoanalytic meanings. For example, in some contexts the response "eyes looking at me"—in some of the cards—is

⁷ Beck, *op. cit.*

given the obvious interpretation of "paranoid reaction." "Puppets" or "marionettes" perceived in a card are interpreted at times to suggest schizoid tendency, as a feeling of being influenced and directed by hostile persons.

Originals and Populars. The fourth scoring category is *originality*—also known as "popularity-originality." This has to do with the rating of a response as one that is commonly given (popular) or as one that is uncommon (original). Investigators and interpreters of Rorschach test responses are not entirely agreed as to which responses shall be scored popular and which original, although there are, of course, many responses about which there is no doubt. However, if users of the Rorschach test are to achieve a satisfactory and essential degree of uniformity in regard to the significance of their results, this problem of popularity-originality will have to be resolved statistically, in a manner similar to that used in other types of tests. Some normative studies have been made in this area.⁸

Interpretation. Scoring of the responses according to the foregoing categories, or according to any one of the modifications and elaborations thereof,⁹ is not an end in itself. The major purpose of the test is to assess the subject's general adjustment and to learn whether he is experiencing psychological difficulties—in short, to get insights into his personality, such as could not be obtained ordinarily by direct questioning.

Although considerable experience under supervision is necessary to learn the techniques of administering and scoring the Rorschach, much more experience and expertness are required for the interpretation of scores as organized meaningful whole. With the Rorschach, the two aspects—a scoring system and skill in (perhaps the art of) psychological interpretation—are essential. The particular items in the responses of the subject are not in themselves most important; the inferences drawn from them are.

⁸ For example: M. R. Hertz, *Frequency Tables to be Used in Scoring Responses to the Rorschach Ink-Blot Test*, Department of Psychology, Western Reserve University, 1946; W. N. Thetford et al., "Developmental Aspects of Personality Structure in Normal Children," *Journal of Projective Techniques*, Vol. 15, 1951, pp. 58-78; R. Carlson, "A Normative Study of Rorschach Responses of Eight-Year-Old Children," *ibid.*, Vol. 16, 1952, pp. 56-65.

⁹ For example, R. L. Munroe, "Prediction of the Adjustment and Academic Performance of College Students by a Modification of the Rorschach Method," *Applied Psychology Monographs*, No. 7, 1945.

A simple example will show how a response is scored for the several categories. The subject states that the inkblot on the first card looks like a bat. This response is interpreted as showing: that the whole blot (W) was perceived (location, or area); that the perception was based upon form, and that the blot is sufficiently similar to a bat in shape so that the response is acceptable (F); that the response is classified under "animal" (content, A); and that the response is commonly given, or popular (P). Hence, using symbols, this response would be recorded as W, F, A, P.

Interrelationships. Having scored each response, tabulated the results, and drawn a psychogram, the next and very significant step is to analyze the *relationships* existing between the frequencies in several categories, since the frequencies in isolation are not all-important. For example, the following are among the relationships investigated: the proportion of responses in each of the scoring categories, this being of great importance; the ratio of wholes to larger details and to minute details; extent to which form is used with other determinants, such as color and shading, form being dominant; extent to which color is used with other determinants, color being dominant; relationship between color and movement responses; frequency with which color is named apart from other determinants; the ratio of human-movement responses to animal-movement responses; ratio of pure movement responses apart from other determinants to movement plus other elements, especially form; percent of original responses. These and other comparisons and interrelationships show the organization and patterned characteristics of the individual's personality, as measured by the Rorschach test. For instance, responses indicative of strong uncontrolled emotion, and thus a sign of possible danger, in one record may be regarded as relatively serious; in another record if there are balancing factors they may be indicative of satisfactory adjustment.

The frequencies of responses in each of the categories and the various ratios and interrelationships are the bases upon which the subject's personality is delineated and described. Without going into great detail, a few illustrations will be given so that the reader may see how responses are interpreted. At the risk of tedious repetition it should be stated again that no single sign is the basis of a diagnosis.

Location. The *location* responses are used mainly in evaluating intellectual aspects of personality: the manner of approach to a per-

ceptual problem; the preferred mode of apperception. A predominance or large percentage of whole responses (W)—that is, emphasis upon the whole rather than the particular—is characteristic of persons of higher levels of capacity for intellectual organization and abstraction. Not only the number of wholes is important here; the originality and appropriateness of the responses must be considered; for simple, popular wholes indicate superficiality and commonplace thinking.

Predominance of common responses of detail (D) is regarded as evidence of concrete, unoriginal, practical mental processes. Responses of unusual detail (Dd) indicate perception of the unusual, associated at times with precise and critical mentalities. But carried to an extreme, rare detail responses may indicate an obsessive preoccupation with the trivial, often accompanying states of anxiety. Some clinicians have claimed more subtle meanings in location responses; as, for example, that a marked concern with edge details signifies the presence of an escape mechanism.

Rigidity of approach to problems is said to be indicated by a subject who uses the same procedures with all cards in making his reports, beginning with wholes, then proceeding to larger details, minute details, and white spaces. The reverse of this order is found less frequently. Rigidity is distinguished from *orderliness* of approach, the latter being indicated when a uniform procedure is followed in most location responses, but when, also, variations are used on occasion. Psychotic and some emotionally unstable persons are confused and chaotic in their approach, no order or plan being observable.

Form. If a person's *form perception* is clear and accurate (F+ or F), he is said to have firm control over his intellectual processes and behavior. By contrast, the schizophrenic, whose behavior is disorganized and whose perceptions are distorted, often reports inappropriate and bizarre forms for the blots (F-). A high percentage of form scores independent of other determinants is regarded as evidence of restricted emotional and social adjustment (repression or suppression). Form scores combined with other determinants indicate the degree of intellectual activity.

Color. The subject's responses to color in the blots provide the most direct evidence of the subject's impulsive life and emotional relationships to his environment, the combination of color and form and extreme accent on color alone being regarded as the most significant signs. Color usually arouses some degree of feeling or emotion in the

respondent. The subject may entirely suppress his response; he may spontaneously respond solely to the color without associating it with a form or object; he may combine color and form, one or the other being dominant. The degree to which color is the sole or principal determinant in a subject's responses is viewed as indicative of his emotional intensity. The extent and manner in which he combines color with form is evidence of the degree of control over emotional impulse. Predominance of form-color responses (FC), perceived as meaningful wholes, imply optimal emotional control, accompanied by a capacity for social adaptability. Color-form responses (CF), in which form is secondary, when dominant suggest a somewhat impulsive, egocentric personality. A large proportion of pure color responses (C) is indicative of emotional impulsiveness.

An additional kind of color response, "color shock," has been observed. This happens when the respondent is thrown off balance by the presentation of a colored card, especially when it follows a black-and-white blot. Shock is shown by a delay in reaction time, by exclamations, by peculiar responses; or by inability to respond to color. Color shock is believed to imply anxiety neurosis; the person's ability to respond is seriously impaired by loss of emotional equilibrium due to affect produced by color. Color shock has not only been found frequently with neurotic persons; it seems to be typical of persons more seriously disturbed mentally.

The concept of "color shock" and actual appearance of the phenomenon, however, have been questioned and subjected to experimental study. Most of the experimental data do not support the original concept.¹⁰ But critics of these statistical studies maintain that the concept of "color shock" and interpretation of color responses have been handled in a mechanical manner, without due regard to the individualized way in which color is handled in the record. They hold that *how* the subject responds to color, if at all, is more significant than statistical counts, in judging the impact of color stimuli upon emotional responsiveness. These critics of the usual statistical analyses point out that the following aspects of response must be evaluated in clinical interpretations of color responses: color selection, color shy-

¹⁰ See, for example, J. A. Perlman, "Color and the Validity of the Rorschach 8-9-10 Percent," *Journal of Consulting Psychology*, Vol. 15, 1951, pp. 122-126; B. T. Meyer, "An Investigation of Color Shock in the Rorschach Test," *Journal of Clinical Psychology*, Vol. 7, 1951, pp. 367-370.

ness, color denial, color avoidance, and color disregard with objective disturbance.¹¹

The disagreement noted regarding this area of Rorschach interpretation exemplifies, in general, a principal basis of difficulty encountered in efforts to objectify interpretations of responses and to validate the instrument as a whole: namely, the conflict between those who would break down the responses into elements for purposes of statistical analysis, and those who maintain that by so doing the significance of each element is destroyed, since each must be viewed in its relationships to the whole. This latter group maintain, also, that available statistical methods are inadequate for handling this problem of primary importance.

Shading. Responses that use shading are said to be related to the manner in which the subject meets his affectional needs. These responses are interpreted as being related to anxiety, depressed attitudes, and feelings of inadequacy.

Movement. The *movement* score (especially human movement), according to Rorschach, is evidence of the richness of one's associative life; the higher the score, the richer the associations and the richer the imaginative life. A large number of reports of human movement, combined with responses to color but not outweighed by them, indicates superior creativeness and giftedness. A high frequency of human movement with little or no color response is said to be characteristic of persons having a rich inner life but little affective response to the outside world. Such persons were named *introversive* by Rorschach. On the other hand, much color response, accompanied by few reports of human movement, indicates the *extraversive* personality.

Beck, one of the productive students of the Rorschach, reports that the significance of movement responses differs with various personality organizations.¹² In emotionally stable adults and in some neurotics, it is as stated above. In cases of schizophrenia, movement response is indicative of a highly subjective and personal experience. In adults having adjustment problems but without psychosis, it represents fantasy living; in the manic, it indicates egocentric wish fulfillment.

Klopfer, who has published extensively on the Rorschach, makes

¹¹ See Klopfer et al., *op. cit.*, pp. 338 ff.

¹² S. J. Beck, "Psychological Processes in Rorschach Finding," *Journal of Abnormal and Social Psychology*, Vol. 31, 1937, pp. 482-488.

a distinction between reports of human movement and those of animal movement.¹³ In respect to the former, he agrees essentially with Rorschach's interpretation. But a large proportion of the latter, according to Klopfer, indicates that the person is functioning on a "level of instinctive prompting" rather than at the level of creative activity.

Hertz and Kennedy have added still another interpretation to movement responses.¹⁴ They report that a high movement score combined with satisfactory form, originality, detail, and organization indicates superior intellectual ability.¹⁵

Content. The number, proportions, and kinds of things represented in the responses have been variously interpreted as having psychoanalytic significance (fantasy life, symbolic meaning); as showing the amount of stereotyped thinking in the subject; as a sign of maturity or immaturity; as revealing feelings of inferiority; as reflecting the subject's interests, obsessions, and compulsions. Much of this content interpretation is tentative and needs experimental confirmation.

Originals and Populars. The percentages and number of *original* and *popular* responses, as might be expected, are taken as evidence of the subject's level of intelligence, though the quality of the originals must be evaluated; for these might be no more than bizarre items or distortions of perception.

Interrelationships. The relationships *within* each of the scoring categories and *among* the several categories provide the materials from which an individual's personality structure and organization are inferred. Various clinicians and investigators have arrived at norms and response proportions, for which they have offered formulas to be used in determining the characteristics of an individual's personality, such as intellectual aspects, emotional control, insecurity and anxiety, introversive, extraversive, or ambiequal "types," etc.¹⁶

From the foregoing sketch of the possible meanings of various scores it should be clear that no single indicator is relied upon ex-

¹³ B. Klopfer and D. M. Kelley, *op. cit.*

¹⁴ M. R. Hertz and S. Kennedy, "The M Factor in Estimating Intelligence," *Rorschach Research Exchange*, Vol. 4, 1940, pp. 105-106.

¹⁵ Some of the recent experimental work on movement has been concerned with its relation to the perceptual process. See R. Arnheim, "Perceptual and Aesthetic Aspects of the Movement Response," *Journal of Personality*, Vol. 19, 1951, pp. 265-281; G. S. Klein and H. J. Schlesinger, "Perceptual Attitudes toward Instability: Prediction of Apparent Movement Experiences from Rorschach Responses," *ibid.*, pp. 289-302.

¹⁶ For several of these, see Bell, *op. cit.*, Chapter 16.

clusively in determining the nature of a given personality and that the major task of the clinician is to interrelate the indicators in such a way as to yield a meaningful whole. It is obvious, also, that scoring techniques among the experts are as yet not uniform, being in a state of development. It should be apparent, too, that there is a high degree of subjectivity in both the scoring and interpretation of responses—and necessarily so in dealing with an unstructured test.

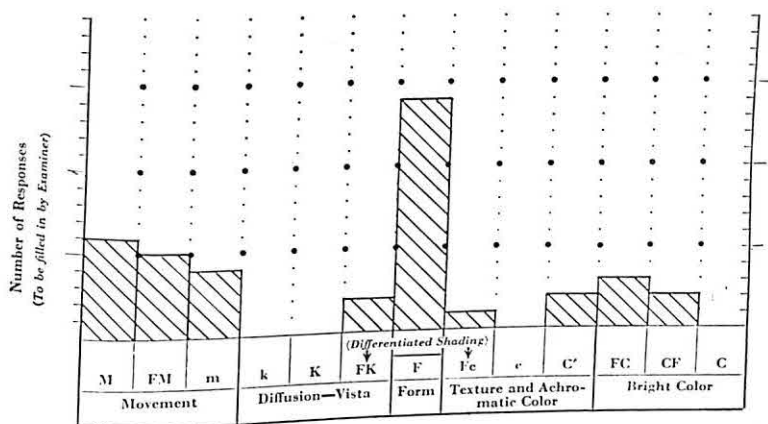


FIG. 19.2. Rorschach response profile. (From the *Individual Record Blank*, by B. Klopfer and H. H. Davidson, New York: World Book Co. By permission.)

This means, of course, that skill in the use of such instruments and the attainment of maximum validity can be achieved only after carefully supervised practice and experience.

Figure 19.2 is a profile form devised by Bruno Klopfer and Helen H. Davidson.¹⁷ As reproduced here, it shows the number of responses of an actual case in each category, to illustrate the manner in which results are portrayed. Table 50, also by Klopfer and Davidson, is shown so that the student may see more clearly the extent to which the various scores are interrelated and comprehend more clearly what is meant by interpretation of Rorschach responses as a whole. The information in this figure and the table provide the starting materials for interpretation.

Personality "Structure." The Rorschach test is a "multidimensional instrument" that is intended to yield information regarding the "struc-

¹⁷ Published by World Book Co. (By permission.)

ture" of the subject's personality. Three major "dimensions" are appraised: namely, conscious intellectual activity, externalized emotions, and internalized emotions. Each of these is "measured" in terms of the several scoring categories already explained. By "structure" is meant the manner in which the various personality aspects or traits are interrelated so as to produce each of the three major "dimensions"

RELATIONSHIPS AMONG FACTORS

Total Responses (R) =

Total Time (T) =

Average time per response ($\frac{T}{R}$) =

Average reaction time for Cards I, IV, V, VI, VII =

Average reaction time for Cards II, III, VIII, IX, X =

 $\frac{\text{Total F}}{R} = F\%$ $\frac{FK + F + Fc}{R} = \%$ $\frac{A + Ad}{R} = A\%$

Number of P =

Number of O =

 $(H + A) : (Hd + Ad) = :$ $\text{sum C} = \frac{FC + 2 CF + 3 C}{2} =$

M: sum C = :

 $(FM + m) : (Fc + c + C') = :$ $\frac{\text{No. of responses to Cards VIII, IX, X}}{R} = \%$

W: M =

Succession:

Rigid Orderly Loose Confused

(Place a check mark at the appropriate point on the scale above)

Estimate of Intellectual Level

Intellectual Capacity

... Very Superior

... Superior

... High Average

... Low Average

... Dull Normal

... Feeble-minded

Intellectual Efficiency

... Very Superior

... Superior

... High Average

... Low Average

... Dull Normal

... Feeble-minded

Note that this estimate is based mainly on the following:
 number and quality of W
 number and quality of M
 level of form accuracy
 number and quality of O
 variety of content
 succession

Manner of Approach

W(____%) D(____%) d(____%) Dd and S(____%)

Enter the location percentages in the spaces above. Compare these percentages with the norms shown in the box below, by placing a check mark opposite the appropriate range of percentages.

W	D	d	Dd and S
< 10% ((W))	< 30% ((D))		
10-20 (W)	30-45 (D)		
20-30 W	45-55 D	5-15 d	< 10% Dd S
30-45 W	55-65 D	15-25 d	10-15 Dd S
45-60 W	65-80 D	25-35 d	15-20 Dd S
> 60 W	> 80 D	35-45 d	20-25 Dd S
		> 45 d	> 25 Dd S

[4]

TABLE 50. Outline for analyzing Rorschach responses.¹⁸

which, themselves, are interrelated so as to produce the individual's whole personality. The organization, or "structure," of a person's traits is said to indicate how he experiences life around him and how he utilizes some of his experiences. What are his characteristic perceptions, attitudes, and behaviors that result from the organization of his traits? For example, in regard to "structure," some of the questions that the Rorschach pattern of responses are intended to answer are: Is the subject's mentality original or stereotyped? Are his abilities creative or reproductive? Is he less adaptable or more adaptable to reality? Is the "inner" or the "outward" life stronger? Or, more spe-

¹⁸ From the *Individual Record Blank*, by B. Klopfer and H. H. Davidson, New York: World Book Co. (By permission.)

cifically, how and to what degree does the individual control his emotions and feelings? Is he prone to anxiety?

Taking one of the three major "dimensions"—intellectual activity—as an example, it would be appraised by the scores on the following aspects of the responses. Each aspect is accompanied by the factors regarded as significant elements in interpretation of mental level.

Perception of form: quality and percentage of total responses; clearness vs. vagueness; accuracy vs. inaccuracy (F, F+, or F-).

Perception of wholes: number and percentage of total responses, ability to integrate; ability in abstract and theoretical activity.

Perception of major and minor details: number and percentage of total responses; concrete or practical intelligence.

Organization: ability to perceive and create new wholes; creative ability.

Original and popular responses: qualitative richness of response vs. commonplace, stereotyped thinking.

Animal content: "sterility" or immaturity of thought processes (an excess of animal responses) vs. rich and insightful thought processes (perception of forms of a variety of categories).

Productivity: total number of responses; intellectual energy.

Sequence: rigid, or orderly and flexible, or loose, or confused approach to a problem, indicating level of intellectual control.

Movement (mainly human movement): inventive and creative abilities vs. stereotyped thinking.

Reliability. As with all other types of psychological tests, reliability and validity of the Rorschach need to be demonstrated if this instrument is to attain maximum usefulness. It is undoubtedly true that widely experienced clinicians using this test are able to draw some remarkably acute inferences from Rorschach protocols; but this is a highly subjective procedure. Ideally, it is desirable to demonstrate objectively the Rorschach's reliability and validity; and considerable research effort—both clinical and experimental—is being devoted to that end.

There are, however, several problems and principles that must be kept in mind in evaluating reliability and validity studies on the Rorschach. First, interpretation of responses is "global"; that is, the significance of responses is determined by the total pattern, whereas analysis into parts and part-scores for the purpose of statistical treatment does violence to the principle underlying the test and to the meanings of the responses. As yet there are no statistical methods for

dealing with integrated, "global," patterns of responses. The several methods thus far suggested are quite inadequate for handling this complex problem. Second, rating or scoring of responses introduces a greater subjective element (than in the case of intelligence tests) because there are no right or wrong answers, the variety of responses is considerable, and normative data are meagre. As criteria of validity, differential clinical diagnoses are not sufficiently uniform or reliable, and descriptions of behavior are non-quantitative and subjective to an appreciable degree.

The foregoing points emphasize the unusual difficulties encountered in attempts to validate and objectify projective techniques in the same manner as other types of instruments have been. Although the problems of validity and reliability have not been solved, much work has been done. We shall describe briefly the principal methods employed, together with the present status of the findings.

Studies devised to show degree of Rorschach reliability may be classified under five types, as shown below. Each approach will be briefly explained and evaluated.

1. Parallel series of cards
2. Split-half correlations
3. Test-retest correlations
4. Matching interpretations of Rorschach records
5. Attempts on the part of the subject to falsify responses and misrepresent themselves

Parallel Series. The use of two *parallel series of cards* is intended to be analogous to the use of two equivalent forms of a test of intelligence or of specific aptitude. In this country a set of parallel cards has been devised by Harrower and Steiner.¹⁹ Data in the manual show that this newer set of cards elicit much the same types of responses as do the Rorschach cards, and in similar proportions. Unfortunately, however, no systematic statistical analyses of scores are presented in the manual; nor has there been sufficient evidence from other sources to warrant any but the most general conclusions regarding the degree of equivalence of the two sets of cards. The original Rorschach and the Harrower-Steiner have yielded sufficiently similar results for the

¹⁹ M. R. Harrower and M. E. Steiner, *Psychodiagnostic Inkblots*, New York: Greene and Stratton, 1945.

same subjects as to warrant the use of the latter as a supplemental instrument.

It should be pointed out that devising two equivalent or comparable sets of inkblots is a much more difficult task than devising two sets of equivalent questions or problems; for unless two inkblots are identical, their stimulus values will be different in some degree. For this reason, the newer sets of cards are appropriately called "parallel" rather than equivalent. The more precise evaluation of the parallel forms, and reliability studies based on them, must await detailed quantitative data and intercomparisons.

Split-half. When the *split-half* method is used to estimate reliability, responses to odd-numbered cards are correlated with those given to even-numbered cards. That is, the score in each of the categories obtained on odd-numbered cards is correlated with the corresponding score on the even-numbered (e.g., number of W's, F's, D's, etc.).²⁰ This method has yielded reliability coefficients ranging from low to high: from approximately .60 to .90.

For several reasons, the *split-half* method has been virtually abandoned in Rorschach studies. First, there is the usual argument that the significance of a Rorschach protocol is to be found in the whole, integrated set of responses; therefore, working with isolated variables is invalid. This argument does not have much merit, since isolation of a variable for statistical analysis is one matter, while isolation for interpretation is quite another. The second and more pertinent point is that there are marked variations in the stimulus value of each of the ten cards; hence, since they are not similar in functions or behaviors to be elicited, the odd-even method should not be employed. A third point—one that applies to all quantitative analyses of the Rorschach—is that scoring of responses introduces a strong subjective factor, due to lack of normative data and to the exceedingly large number and variety of responses possible.

Test-retest. The *test-retest* method also has yielded coefficients

²⁰ Cf. G. R. Thornton and J. P. Guilford, "The Reliability and Meaning of Erlebnistypus Scores in the Rorschach Test," *Journal of Abnormal and Social Psychology*, Vol. 31, 1936, pp. 324-330; P. E. Vernon, "The Rorschach Inkblot Test," *British Journal of Medical Psychology*, Vol. 13, 1933, pp. 89-118, 271-295; M. R. Hertz, "The Reliability of the Rorschach Inkblot Test," *Journal of Applied Psychology*, Vol. 18, 1934, pp. 461-477; M. R. Hertz, "Rorschach: Twenty Years After," *Psychological Bulletin*, Vol. 39, 1942, pp. 529-572.

from low to high, the maximum being about .90. The magnitude of the coefficient has varied roughly with the length of interval between tests.²¹ Swift, for example, retested preschool children at several different intervals, with the following results:²²

After 14 days, coefficients varied from .59 to .83 for the several scoring categories.

After a 10-month interval, coefficients varied from .18 to .53 for the several scoring categories.

The findings for the 14-day interval are much more significant as evidence of reliability (internal consistency) of a projective test, especially with very young children who are developing rapidly and in whose lives an interval of ten months is a relatively large and significant period.

Another reliability study, using twenty chronic schizophrenics, yielded quite favorable results.²³ Reliability coefficients were:

Location categories:	.50 (D) to .95 (W)
Determinants:	.16 (FC) to .96 (k)
Content:	.36 (Ab) to .94 (Hd)
Relative scores:	-.17 (8-9-10%) to .84 (W%, d%, F+%)

On the whole, with a few exceptions, the coefficients ranged from moderate to quite high. Low correlation coefficients, the authors state, were due to infrequencies of responses (low N) and narrow ranges of N within the categories concerned. The data suggest, also, that, with one exception, there were no significant differences between means. An additional reliability test was applied: test and retest tabulation sheets were matched by two experienced Rorschach exam-

²¹ Cf. Z. Piotrowski, "The Reliability of Rorschach's Erlebnistypus," *Journal of Abnormal and Social Psychology*, Vol. 32, 1937, pp. 439-445; H. W. Brosin and E. O. Fromm, "Some Principles of Gestalt Psychology in the Rorschach Experiment," *Rorschach Research Exchange*, Vol. 6, 1942, pp. 1-15; A. I. Rabin and M. H. Sanderson, "An Experimental Inquiry into Some Rorschach Procedures," *Journal of Clinical Psychology*, Vol. 3, 1947, pp. 216-225; I. A. Fosberg, "An Experimental Study of the Reliability of the Rorschach Psychodiagnostic Technique," *Rorschach Research Exchange*, Vol. 5, 1941, pp. 72-84.

²² J. W. Swift, "Reliability of Rorschach Scoring Categories with Preschool Children," *Child Development*, Vol. 15, 1944, pp. 207-216.

²³ J. D. Holzberg and M. Wexler, "Predictability of Schizophrenic Performance on the Rorschach Test," *Journal of Consulting Psychology*, Vol. 14, 1950, pp. 395-399.

iners. They were correct in 85 percent of their selections, this being significant at the one-percent level.

The test-retest method should be used with caution in connection with projective methods, for these reasons: There may be some carry-over resulting from mere recall of previous responses. Since the scoring of determinants is dependent, in part, upon the subjects' introspection during the inquiry, another possibility of inconsistency is introduced in retesting. Is the subject always able to state exactly which determinants (elements) were primary or involved in his perceptions? If the interval between examinations is a significant one, some changes in behavior may have taken place due to maturation, therapy, or environmental factors.

It appears that Rorschach results are more stable for some classes of persons and for some age groups than for others. The question to be asked, therefore, is one of *specific* reliabilities: reliable for whom and under which conditions? Partial answers have been given to this question; but much more research remains to be done.

Matching. As a method of estimating Rorschach reliability, matching is regarded by many experts as most satisfactory, since the whole Rorschach picture is thereby kept intact. This is a reliability technique in the sense that it attempts to answer the question: "Do the responses have the same meanings for different experts?" Internal consistency of responses, then, is estimated in terms of consistency of meaning conveyed by the responses; that is reliability of scoring and of interpretations. Krugman reports, for example, that three judges agreed perfectly in matching twenty Rorschach records with interpretations that had been made by others. Also, in this same study it was found that when two independent interpretations of the twenty response records were made, the two agreed essentially in respect to the significance of about ninety percent of the response data and showed partial agreement on ten percent. In another instance when twenty-five records were matched with interpretations, using six judges, the average coefficient of contingency was .87.²⁴

²⁴ J. I. Krugman, "A Clinical Validation of the Rorschach with Problem Children," *Rorschach Research Exchange*, Vol. 6, 1942, pp. 61-70. Actual percentages were: essential agreement, 89.6%; partial agreement, 10%; disagreement, 0.4%. The coefficient of contingency is an index of correlation which is used when both variables are classified in two or more categories rather than being given quantitatively in the form of a continuous or a discrete variable. See also J. O. Palmer, "A Dual Approach to Rorschach Validation: A Methodological Study," *Psychological Monographs*, No. 325, 1951.

Attempts to Mislead. The remaining approach—attempts to falsify responses—appears to be neither an appropriate nor a reasonable method of estimating Rorschach reliability. Some psychologists have called this method “testing the limits of reliability.” The results on almost any type of testing device can be more or less distorted by testees, depending upon their intelligence and degree of psychological sophistication. When this method is used, the experimenter is not actually trying to add evidence to the problem of reliability. He is seeking an answer to the question, “Is the Rorschach test resistant to a person’s attempts to falsify responses and to misrepresent his personality?” The answer is that some changes in the patterns of response can be effected through deliberate effort; but the amounts and directions of change depend upon several factors: namely, relative personality maturity and versatility to begin with, normality versus maladjustment, intellectual level, amount of knowledge about and definiteness of “set” toward the test.²⁵

At the same time, there is evidence to indicate that experimental subjects who are simply instructed to try to make a good or poor impression do not succeed, in a large percentage of cases, in altering their response patterns. In one study of twenty persons, using the several scoring categories, only three subjects succeeded to an appreciable extent (5-percent level of significance) and only two other subjects to a moderate degree (10-percent level).²⁶

Summary. In summary, it appears, on the whole, that the functions examined by the Rorschach method have a fair degree of reliability in terms of internal consistency and retest after a relatively brief interval. When scoring categories and elements within them are isolated, some functions are found to be more stable than others. The instrument’s reliability appears to be greater when the “global” or patterned character of response records are studied than when scoring categories are isolated. In evaluating the findings of reliability studies on the Rorschach, it is essential to bear in mind the complexity of the

²⁵ For example, M. L. Hutt et al., “The Effect of Varied Experimental Sets on Rorschach Test Performance,” *Journal of Projective Techniques*, Vol. 14, 1950, pp. 181–186; R. G. Gibby, “Stability of Certain Rorschach Variables under Conditions of Experimentally Induced Sets: The Intellectual Variables,” *ibid.*, Vol. 15, 1951, pp. 3–26.

²⁶ A. L. Carp and A. R. Shavzin, “The Susceptibility to Falsification of the Rorschach Psychodiagnostic Technique,” *Journal of Consulting Psychology*, Vol. 14, 1950, pp. 230–233.

instrument, the subtleties of interpretation, and the influence of extraneous factors (e.g., "set," motivation, verbal handicaps, previous experience, attitude toward examiner) upon perception. An individual's responses in a test situation are not the product of only his "constant" personality traits reacting to constant test materials.

From the data at present available, we must conclude that reliability of the Rorschach test has not been unequivocally demonstrated *when the usual methods and standards are used*, but that the evidence on the whole is more favorable than unfavorable. It may be, however, that the usual conception and criteria of reliability (such as applied to tests of intelligence and specific aptitudes) are not and should not be applicable to an instrument like the Rorschach, which is unstructured and thus permits variations in responses, which is not readily quantified, and which has been employed largely to test and describe personalities who are maladjusted or in a fluid state. Perhaps the more significant result is the very high percentage of agreement of judges in the interpretation of responses. This criterion, however, does not satisfy the standard definitions of reliability. In regard to the Rorschach at present, we may say that some researches indicate satisfactory reliability, as usually defined, while others do not; but there is a high degree of consistency among Rorschach examiners regarding interpretation of response data.

Validity. It is entirely reasonable to ask how we know that the stated relationships exist between the scoring categories and the personality traits they are said to indicate. How is it known that color responses, for example, have the emotional significance attributed to them, or that movement responses are related to richness of associations and imaginativeness? The answers to these and similar questions are to be found in the fact that Rorschach himself devised and used the inkblots clinically in the diagnosis of personality and experimentally with persons of known personality traits. In so doing he found the response differences reported. Since the publication of his test in 1921, his disciples have also attempted to establish its validity by similar means.

Known Groups. There are several types of validating studies.²⁷ The first method, used by Rorschach, is the intercomparison of groups of known differences in intelligence and personality organization. For

²⁷ Cf. M. R. Hertz, "Validity of the Rorschach Method," *American Journal of Orthopsychiatry*, Vol. 11, 1941, pp. 512-519.

the most part Rorschach used mental patients who presented rather clearly discernible extremes of certain traits. In addition he tested artists, scholars, persons of average abilities, and some mental deficient. From group to group he and other investigators found significant differences in responses and in characteristic patterns. For example, in the neurosis pattern, the following distinguishing signs, among others, are some of the most important reported: very few movement responses, color shock, shading shock, few or no form-color responses, noncombined form responses constituting fifty percent or more of the total number of responses, refusals to one or more cards, small total number of responses. This pattern is interpreted as signifying lack of social adaptability, excessively rigid control, suppression of spontaneity and originality, and anxiety. The value of this method of validation depends, in the first place, upon the accuracy of the clinical diagnoses of the subjects whose Rorschach records are compared with their clinical classification (based upon their case histories). Results thus obtained from numerous clinical reports are considered reasonably satisfactory.²⁸

Since Rorschach's original work, many and varied validating methods have been used. These may be classified as follows:

1. Rorschach diagnoses compared with diagnoses by psychotherapists and clinical interviewers
2. Rorschach findings compared with consistent observation of behavior over an adequate period of time
3. Matching Rorschach interpretations with clinical case reports
4. Comparing Rorschach protocols obtained before and after therapy
5. Single Rorschach variables, or a combination of a few, related to observed aspects of behavior
6. Experimental validation: (a) influencing the subjects; (b) varying the stimulus; (c) relation of Rorschach responses to physiological activity

Each method will be briefly described and general results briefly noted.

²⁸ A detailed summary and a bibliography are in Bell, *op. cit.*, pp. 138-149. More recent selected references are included in subsequent footnotes in this chapter.

Comparisons of Diagnoses. Individuals are examined and diagnosed with the Rorschach. Also, another staff member makes a psychiatric diagnosis after interviews. Findings are compared. In one study of 26 children referred to a clinic, the two sets of diagnoses agreed in 62 percent of the cases before psychotherapy. A year later, Rorschach and psychiatric diagnoses agreed in 89 percent of the cases, all the shifts having been made in the psychiatric classification.²⁹ Although not all reports are as favorable as this one, the results in general are reasonably satisfactory—sufficiently so that the Rorschach is widely used in clinics as a diagnostic aid.

Comparisons with Observations. Behavioral observations of selected individuals are made continuously over a period of time by qualified persons. The subjects observed are usually maladjusted children or adults diagnosed as pathological, or non-clinical school children. These observations are compared with Rorschach findings in respect to certain aspects of personality, such as intellectual functioning, anxiety, emotional expression, etc. The observations may be made in a camp for children, a community or recreation center, in a school, and the like. The results of studies of this type have provided only partial validation.³⁰ In one of the more favorable reports in this area of validation teachers' descriptions and ratings of personalities of thirty children (non-clinical cases) were matched with the Rorschach responses of these subjects. The correspondence found was at the one-percent level of significance (which means that the relationship would occur by chance only once in a hundred instances).

Matching. The most frequently used method of studying validity involves comparisons of Rorschach reports with clinical case reports. Some of the studies have used deviant personalities; others have used

²⁹ M. G. Siegel, "The Diagnostic and Prognostic Validity of the Rorschach Test in a Child Guidance Clinic," *American Journal of Orthopsychiatry*, Vol. 18, 1948, pp. 119-133. Also J. D. Benjamin and F. G. Ebaugh, "The Diagnostic Validity of the Rorschach Test," *American Journal of Psychiatry*, Vol. 94, 1938, pp. 1163-1178; M. R. Hertz and B. B. Rubinstein, "A Comparison of Three 'Blind' Rorschach Analyses," *ibid.*, Vol. 9, 1939, pp. 295-314.

³⁰ R. A. Young and S. A. Higginbotham, "Behavior Checks on the Rorschach Method," *American Journal of Orthopsychiatry*, Vol. 12, 1942, pp. 87-94; J. L. Singer and H. E. Spohn, "Some Behavioral Correlates of Rorschach's Experience-Type," *Journal of Consulting Psychology*, Vol. 18, 1954, pp. 1-9; J. W. Swift, "Matchings of Teachers' Descriptions and Rorschach Analyses of Preschool Children," *Child Development*, Vol. 15, 1944, pp. 217-224.

clinical groups; others have equated two groups, except for one important factor, the significance of which was to be tested (e.g., adolescents in foster-homes compared with those in institutions).

An outstanding study in this area is Krugman's.³¹ Rorschach interpretations of 25 problem children were matched by 5 judges with clinical case reports. For these 125 pairings, a contingency coefficient of .85 was found. For the 5 judges, the average of correct matchings was 84 percent. The next step was to compare Rorschach interpretations with case records in regard to each of the following: (1) intellectual aspects; (2) emotional aspects; (3) diagnosis; (4) total personality. Essential agreement was found in 73 percent of the comparisons, fair agreement in 21 percent, and slight agreement in the remainder. These are highly favorable findings.

Changes after Treatment. When Rorschach reports obtained before and after various forms of treatment are compared, as a method of validating the test, the hypothesis is that the test record should reflect the personality changes that have taken place in the interim as a result of therapy. Two studies—one involving psychoanalysis and the other involving insulin treatment—will be mentioned.³² In the first, 36 persons were involved, for all of whom there were significant changes as between the "before" and the "after" Rorschach records; and these changes were related to the trends reported by the therapist: for example, improved emotional control, improved intellectual functioning. On 14 out-patient subjects, the Rorschach examiner and the therapist were in essential agreement regarding improvement. In the case of 22 hospitalized patients, the two clinicians agreed on 10 in regard to direction and degree of improvement; for the remaining 12, the therapist reported "social recovery" (a nebulous matter), whereas their Rorschach responses showed no improvement.

When insulin treatment was used on schizophrenics, those who

³¹ J. I. Krugman, "A Clinical Validation of the Rorschach with Problem Children," *Rorschach Research Exchange*, Vol. 6, 1942, pp. 61-70. Also, W. Goldfarb, "Rorschach Test Differences between Family-reared, Institution-reared, and Schizophrenic Children," *American Journal of Orthopsychiatry*, Vol. 19, 1949, pp. 624-633.

³² M. J. Rioch, "The Use of the Rorschach Test in the Assessment of Change in Patients under Psychotherapy," *Psychiatry*, Vol. 12, 1949, pp. 427-434; Z. A. Piotrowski, "Rorschach Manifestations of Improvement in Insulin Treated Schizophrenics," *Psychosomatic Medicine*, Vol. 1, 1939, pp. 508-526; C. Windle, "Psychological Tests in Psychopathological Prognosis," *Psychological Bulletin*, Vol. 49, 1952, pp. 451-482.

showed behavioral improvement also showed improved performance on the Rorschach: namely, increase in speed of reaction and of response, improved verbal form and logical content, clearer perceptions and more relevant responses, and improved emotional control.

Separate Variables Related to Behavior. Without going into the details and tentative findings in this area of validation, we shall indicate the kinds of studies undertaken and the general conclusions. This is called the "molecular" approach because, instead of making a "molar" or "global" interpretation, only one or a few Rorschach variables are studied in relationship to selected aspects of behavior. The Rorschach variables most frequently investigated have been color, movement, and form responses as they are related to certain aspects of behavior, such as introversion-extraversion (as measured by a personality inventory), intellectual capacity (using a test of intelligence).³³ On the whole, the results in this area have been inconclusive, since the functions being measured by the Rorschach and by the criteria are not necessarily identical or similar.

Experimental Validation. This method takes several forms. The subjects used may be influenced through experimentally induced tensions, through hypnosis or drugs, through brain-surgery, or through electric shock treatment. When the Rorschach test is administered "before" and "after," it is possible to estimate the effects of the changed conditions. The general conclusion is that the findings are favorable as indications of validity.³⁴ There are two principal obstacles to the employment of this method: first, the number of subjects is necessarily small in each study, due to the great amount of time required by each case and to the relative difficulty of obtaining subjects; and, second, the paucity of scientific information as to organic changes in subjects treated by means of surgery, drugs, or electric shock.

Experimental studies in which the stimulus has been varied have been concerned almost entirely with the influence of color upon re-

³³ For example, R. W. Gardner, "Impulsivity as Indicated by Rorschach Test Factors," *Journal of Consulting Psychology*, Vol. 15, 1951, pp. 464-468; J. Wishner, "Rorschach Intellectual Indicators in Neurotics," *American Journal of Orthopsychiatry*, Vol. 18, 1948, pp. 265-279.

³⁴ For example, E. Stainbrook, "The Rorschach Description of Immediate Post-convulsive Mental Function," *Character and Personality*, Vol. 12, 1944, pp. 302-322; M. Williams, "An Experimental Study of Intellectual Control Under Stress and Associated Rorschach Factors," *Journal of Consulting Psychology*, Vol. 11, 1947, pp. 21-29.

sponses to the inkblots.³⁵ The purpose of the experiments was to test the stimulus value of color in the cards and the hypothesis of "color shock." Although the results have not been entirely consistent, most researches have raised serious doubts concerning the earlier conceptions of the role of color in this test. In response to these largely negative results, some clinicians point out that, while the reported data are statistically relevant to the isolated stimulus of color, they are not necessarily relevant to the interpretation of color responses in the context of the whole. They emphasize, also, that the statistical findings refer to group trends but do not negate the fact that in some cases the color of the cards has a disruptive influence.³⁶

The third experimental approach is to study the relationships between Rorschach responses and changes in physiological activity. For example: what is the relationship between "color shock" responses and galvanometric response? What are the effects of experimentally induced stress, cardio-vascular activity, and Rorschach responses? Thus far, the relatively few available experimental reports have demonstrated that there is a degree of association between autonomic responsiveness and Rorschach data.³⁷

Summary. Although specific approaches to validation have differed in detail, they are essentially of two broad types: experimental and matching, the latter having been the more widely used and the more significant. The matching methods are based, essentially, upon the principle that descriptions and diagnoses derived from the Rorschach should correspond closely with those obtained by other clinical means—usually those resulting from the long and slow process of clinical interview and case histories. Such approaches are of three kinds, each differing only slightly from the other two: (1) inter-comparisons of Rorschach responses of known groups; (2) "blind" diagnoses of individual cases, followed by comparisons with diagnoses otherwise determined; and (3) direct comparison, or matching, of an

³⁵ For example, J. A. Perlman, "Color and the Validity of the Rorschach 8-9-10 Percent," *Journal of Consulting Psychology*, Vol. 15, 1951, pp. 122-126; H. Sanderson, "Norms for Shock in the Rorschach," *ibid.*, pp. 127-129.

³⁶ For example, E. M. Siipola, "The Influence of Color on Reactions to Ink Blots," *Journal of Personality*, Vol. 18, 1950, pp. 358-382.

³⁷ D. Brower, "The Relation between Certain Rorschach Factors and Cardiovascular Activity before and after Visuo-Motor Conflict," *Journal of General Psychology*, Vol. 37, 1947, pp. 93-95; M. R. Hertz, "Current Problems in Rorschach Theory and Technique," *Journal of Projective Techniques*, Vol. 15, 1951, pp. 307-338.

individual's Rorschach record with extensive clinical study and diagnosis, to determine points of agreement and disagreement. If valid, then, the inkblot test can be used to obtain a diagnosis in much less time, or to supplement or confirm a diagnosis arrived at by other means. Rorschach experts maintain that innumerable investigations support the clinical value of their instrument. They have found, furthermore, that results obtained by the several methods of validation indicate the Rorschach test to be useful in revealing threatening or unwholesome trends in personality development before very serious difficulties actually appear. If this forecasting power of the instrument can be more definitely established, the Rorschach then will become especially valuable for purposes of mental hygiene and preventive psychological treatment. For more definite determination of its forecasting quality, however, an adequate number of longitudinal studies over an appreciable period of the subjects' lives will be necessary. These are as yet not available, for their achievement is beset by many difficulties.

Group Methods. It was to be expected that sooner or later, efforts would be made to devise methods of group administration of Rorschach's individual test, as was the case with intelligence tests. Several techniques have been suggested and are now finding application in several practical fields while, at the same time, being subjected to further experimentation.

In one instance, the Rorschach blots—on slides—are projected on a screen, each for a specified time, the subjects being required to write out their responses. They are later asked to mark the blot reproductions on their blanks and to answer a series of questions so that their responses may be scored according to the standard categories.³⁸

A second method differs from the preceding in that the subjects are provided with a list of responses for each blot from which to choose (multiple choice). A modification of the multiple choice item has been suggested: namely, that the several responses (some "normal" and some "neurotic") be rated by the subjects themselves in order of applicability or likeness to the particular blot to which

³⁸ M. R. Harrower-Erickson, *Large Scale Rorschach Techniques: A Manual for the Group Rorschach and Multiple Choice Test*, Springfield, Ill.: C. C. Thomas, 1945.

they are attached. The sum of the ordinal values assigned to the "neurotic" responses constitutes the score for each blot. Higher scores are, of course, the more favorable since "neurotic" responses are less desirable and should be rated lower in the order of applicability.³⁹

Another group procedure being explored is to have the subjects self-administer the test, following prepared instructions.⁴⁰

From studies thus far reported, it appears that the group methods of Rorschach testing are less successful than the individual method, which permits free, unlimited response, and during which the examiner is able to make observations of the subject's behavior and attitudes. In several fields, however, successful applications of group methods have been reported: in psychiatric screening in the armed services, occupational groupings and selections, differentiating between successful and unsuccessful workers, comparative studies between racial groups in the United States, and in studying personalities of artists and scientists.⁴¹

While group tests can be appropriately and successfully used in studies of *group trends* and *group characteristics*, it is necessary to emphasize again that the individual may be lost sight of in group findings, that a particular individual need not necessarily conform with the trend, and that there is overlapping of scores and characteristics among groups.

As so often happens, however, there are some research data which are not consistent with the rest.⁴² A few studies report only limited value of group Rorschach methods in fields of application such as those mentioned above. Unless it can be demonstrated that the experimental designs and statistical analyses are seriously defective in one group of reports or the other—and this does not appear to be the

³⁹ H. J. Eysenck, "A Comparative Study of Four Screening Tests for Neurotics," *Psychological Bulletin*, Vol. 42, 1945, pp. 659-662.

⁴⁰ R. L. Munroe, "An Experiment with a Self-Administering Form of the Rorschach and Group Administration by Examiners without Training," *Rorschach Research Exchange*, Vol. 10, 1946, pp. 49-59.

⁴¹ For example, M. E. Steiner, *The Psychologist in Industry*, Springfield, Ill.: C. C. Thomas, 1949; A. Roe, "Psychological Examinations of Eminent Biologists," *Journal of Consulting Psychology*, Vol. 13, 1949, pp. 225-246.

⁴² For example, R. G. Anderson, "Rorschach Test Results and Efficiency Ratings of Machinists," *Personnel Psychology*, Vol. 2, 1949, pp. 513-524; A. K. Kurtz, "A Research Test of the Rorschach Test," *ibid.*, Vol. 1, 1948, pp. 41-51.

case—then the problem is to determine the reasons for the discrepancies in results. Why are group methods successful in some situations but not in others? Further detailed research is necessary to answer this question. Such research may reveal not only defects in the group Rorschach, but they may show, for example, that “successful” persons in some occupations may represent a wide variety of personality patterns, or that the reliabilities of the criteria (e.g., foreman’s ratings, ratings by deans of students) are too low. In the meantime, it can be said that group Rorschach methods have been used effectively by some of the specialists in the field.

Evaluation of the Rorschach Test. The Rorschach inkblot method has been shown to have its greatest usefulness in revealing markedly deviant personalities. Its value in differentiating among individuals within the large groups in the middle ground between extremes is limited. This, however, is more or less true of all psychological tests, for two main reasons: differences between individuals within the middle groups are not so pronounced, hence they are more difficult to measure or assess; the instruments are not sufficiently refined to detect the finer differences. There is a third possibility: namely, improved methods of scoring categories, improved interrelationships and sounder interpretations can add to the value of the instrument.

It is pertinent to add, at this point, that often the Rorschach test and Rorschach specialists are asked by critics to demonstrate levels of achievement and of clinical and predictive value such as are not imposed upon non-projective devices, nor, for that matter, upon non-psychological clinical techniques.

Enthusiastic Rorschach exponents recognize that the test is in need of further clinical and experimental research, particularly in respect to scoring, normative studies regarding a variety of factors (e.g., age, sex, cultural and economic status), and longitudinal studies.⁴³

⁴³ To score for maladjustment, R. L. Munroe has suggested the use of a list of indicators of a kind which would give greater objectivity to scores. See “The Inspection Technique: A Method for Rapid Evaluation of the Rorschach Protocol,” *Rorschach Research Exchange*, Vol. 9, 1945, pp. 46-70. For a suggested quantification and standardization of Rorschach scores, see C. Buhler, K. Buhler, and D. W. Lefever, *Development of the Basic Rorschach Score*, with Manual of Directions, 1948 (published by the authors, in mimeographed form, University of Southern California).

Adverse criticism of the Rorschach, from others than its exponents, has not been lacking. Critics have dwelt on its inadequate objectivity, reliance on personal norms, limited validity, restriction to clinical use, and even "cultism." Although these criticisms are warranted to some degree, the fact is that Rorschach exponents themselves have not been unaware of the problems and the unanswered questions; for the literature is replete with their critical publications on the test's reliability, validity, clinical usefulness, guidance value, case studies, and predictive value. With such an approach to the test, it was inevitable that progress in the last decade, both scientific and clinical, should be considerable in the application of the Rorschach to the psychological study of personality and perception and to understanding maladjusted persons.⁴⁴

M. L. Hutt has stated the matter neatly and reasonably: "Clinical practice is, at present, an art. Art, I believe, will always remain an integral part of clinical practice. However, the scientific aspects of clinical work will, in time, become a major portion of this practice as theory, technique and criteria reach a more mature state of psychological development. Meanwhile, I frankly admit the importance of subjective norms, clinical judgment, and subtle influences of intuition. These . . . must be given full play in clinical work in reaching working hypotheses about the [person]. At the same time I rely as much as I can upon all the scientific clues we already have—norms, reliability and validity data, crucial empirical studies, and the like—and utilize these in refining my hunches and delimiting them. Finally, I test these clinical hunches against biographical data, clinical behavior, and all other evidence accumulated about the [person]. In short, I recognize that as a clinician I have two roles to play: the artist and the scientist. I use the former in getting to know the [person] and use the latter to correct my impressions as well as I can. . . . Each role has its place and . . . we must be careful not to confuse the two."⁴⁵

⁴⁴ See the following for a detailed review: J. W. M. Rothney and R. A. Heimann, "Development and Applications of Projective Tests of Personality," Chapter 5 in *Educational and Psychological Testing*, Review of Educational Research, Vol. 23, No. 1, 1953. Also, H. Sargent, "Projective Methods: Their Origins, Theory, and Application in Personality Research," *Psychological Bulletin*, Vol. 42, 1945, pp. 257-293.

⁴⁵ "The Assessment of Individual Personality by Projective Tests: Current Problems," *Journal of Projective Techniques*, Vol. 15, 1951, pp. 389-390. See also F. Halpern, *A Clinical Approach to Children's Rorschachs*, New York: Grune and Stratton, 1953, *passim*.

THEMATIC APPERCEPTION TEST⁴⁶

Description and Procedure. Commonly referred to as the TAT, this projective method consists of thirty pictures plus one blank card. The cards are used in various combinations, depending upon sex and age. Some cards are used with all subjects, while others are used with only one sex- or age-group. The total number of pictures used with any subject is twenty, usually administered in two sessions, ten each time.

The person being examined, according to Murray's instructions, is told that this is a test of imagination, that he is to make up stories to suit himself, and that there is no right and no wrong response. The pictures are shown one at a time, accompanied by simple instructions. The subject is informed that each card shows a scene. He is asked: (1) to tell what he thinks led up to the depicted scene; how it came about; (2) to give an account of what is happening and the feelings of the characters in the picture; (3) to tell what the outcome will be. There are no time limits; in fact, the subject is encouraged to continue for as long as five minutes on a picture. The account should be recorded verbatim if possible.⁴⁷ It is recommended that the testing be followed with an interview to learn the origins of the stories, seeking especially associations to places, names of persons, dates, specific and unusual information. This is an important aspect of the process because it enables the examiner to clarify meanings of stories and more reliably to evaluate their significance, since the subject's accounts are not only a product of his inner personality traits, but may be a more or less superficial reflection of cultural forces (radio, movies, comics, current events, reading materials, etc.). For instance, a girl ten years of age made up an unexpectedly large number of stories dealing with crime and mystery. She had been listening regularly to a radio mystery serial. The fre-

⁴⁶ H. A. Murray, *Thematic Apperception Test Manual*, Cambridge: Harvard University Press, 1943. See also his *Explorations in Personality*, Oxford: Oxford University Press, 1938. This test was introduced in 1935 by C. D. Morgan and H. A. Murray in "A Method for Investigating Fantasies: the Thematic Apperception Test," *Archives of Neurology and Psychiatry*, Vol. 34, 1935, pp. 289-306.

⁴⁷ As in the case of the Rorschach, some specialists in the use of the TAT have introduced their own variations in giving instructions; but basically most follow the formula here given.

quent or compulsive utilization of recent environmental experiences, however, is of significance in interpreting a subject's reports, by virtue of the fact that the person *has* utilized them as representing a conflict on a preconscious level, or as a symbol on an unconscious level.



FIG. 19.3. A Picture from the Thematic Apperception Test. Harvard University Press. (By permission.)

Although pictures are not unstructured to the same degree as an inkblot, those used in the TAT series are sufficiently ambiguous that there is wide latitude for manifestations of individual differences in responses. The TAT is, like the Rorschach, a projective method, but there is a fundamental difference between these two. While the latter is intended to reveal the structure and *organization* (or disorganization) of an individual's personality, the former is devised to bring out primarily the *content* of one's personality: the drives, needs, sentiments, conflicts, complexes, and fantasies. The test is

based upon the principle that when a person interprets an ambiguous social situation, he is apt to reveal aspects of his own personality that he otherwise will not admit, or cannot admit because they are unconscious. The subject, being absorbed in the picture and attempting to construct an appropriate account of it, is "off his guard" and becomes much less aware or quite unaware of himself in the situation. In creating stories based upon equivocal pictures, the individual utilizes and organizes content of his own personal experiences. Everything the subject says is regarded as having meaning. From these stories, the skilled examiner and interpreter draws inferences regarding the subject's personality traits and their organization.

Analysis of Stories. Interpretation of TAT stories may be made in any one of several ways, depending upon the viewpoint of the examiner and the purpose of the testing. But in all instances, the details of an individual's stories must be viewed against known facts of the personality being studied. The stories should not be interpreted *in vacuo*.

In some cases, rereading the stories several times will reveal the subject's basic problems, for repetitive patterns may be found throughout; or it may be found that facts and aspects of different stories constitute a meaningful whole. Or one may make a minute analysis in accordance with the scheme provided by Murray, or in accordance with modifications thereof. We shall briefly outline two schemes of analysis so that the reader may see more clearly the purpose of the TAT. All schemes, however, have this in common: they are intended to disclose personality content, from which the interpreter judges personality organization.

Murray recommends that the content of stories be analyzed into: (1) the forces emanating from the "hero," and (2) the forces emanating from the environment. These two divisions are analyzed under the following six categories.

(1) *The hero*: the character in each picture with whom the subject identifies himself; in whom the subject is most interested; whose point of view, feelings, and motives have been most intimately portrayed. The heroes are to be characterized by the interpreter according to their principal, or idiomatic, traits. (E.g., solitariness, leadership, superiority, criminality, etc.)

(2) *Motives, trends, and feelings of the heroes*: analysis in detail of everything each of the heroes feels, thinks, and does; noting espe-

cially the unusual, the high frequencies, the high and low intensities. Under this category, Murray lists numerous variables (traits) which are scored on a scale from one to five, on the basis of their strength as expressed through intensity, duration, frequency, and importance in the plot. (E.g., abasement, achievement, dominance, conflict, dejection.)

(3) *Forces of the hero's environment*: the general nature and details of the situations, especially human situations, noting especially uniqueness, intensity, and frequency, as well as objects and persons not in the picture but invented by the subject. The environmental forces are then classified according to the effect—realized or promised—upon the hero. Numerous such forces (more than thirty) have been distilled out of the stories. (E.g., rejection, physical injury, dominance, lack, loss.) The strengths of these are rated on a scale of one to five.

(4) *Outcomes*: the comparative strength of the forces emanating from the hero and the strength of those from the environment; the amount of hardship and frustration experienced; relative degrees of success and failure, happy and unhappy endings.

(5) *Themas*: interaction of a hero's needs with environmental forces, together with the successful or unsuccessful outcome for the hero, constitute a simple thema. Combinations or sequences of these are called complex themas. The thema, simple or complex, is actually a synthesis of the elements analyzed under the first four categories, the purpose being to view the several forces in their interrelationships and thus to determine the most prevalent problems, in a given case, arising from internal needs and external forces.

(6) *Interests and sentiments*: choice of topics and manner of dealing with them, displayed especially by the positive and negative value or appeal of various elements in the pictures. (E.g., older women who may be mother figures, older men as father figures, same or opposite sex.)

Tomkins, a former associate of Murray's, has devised his own scheme of analysis. He justifies and differentiates it from Murray's in this way: "Its rationale consists in tapping *varying levels of abstraction* in the hope that significant aspects of diverse types of protocols will be detected by the use of concepts which range from a level of broad generality to a high degree of differentiation." Each story is scored under four main categories: vectors, levels, conditions, qualifiers.⁴⁸

⁴⁸ S. S. Tomkins, *The Thematic Apperception Test*, New York: Grune and Stratton, 1947, Chapter 3. Another and different analytic outline by a former associate of Murray's is by L. B. Bellak, *A Guide to the Interpretation of the Thematic Apperception Test*, Psychological Corporation, 1947. See also, J. I. Lasaga y Travieso and C. Martinez-Arango, "Some Suggestions Concerning the Administration and Interpretation of the TAT," *Journal of Psychology*, Vol. 22, 1946, pp. 117-163.

(1) *Vectors*: the psychological direction of behavior, drives, feelings, etc. The vectors, of which ten have been listed, may have as their objects any thing or person or idea of human interest. Vector means a field of force, or magnitude and direction of force. (E.g., the vector "against," to attack objects; the vector "toward," to approach or enjoy objects.)

(2) *Levels*: the plane of psychological function involved in the story, seventeen being listed. (E.g., object description, intention, wish, night dreams.)

(3) *Conditions*: any psychological, social, or physical state which is not in itself behavior, striving, or wish: that is, conditional qualities of behavior. Two major divisions have been made: (a) states with negative factors or forces—called valences: such as lacks, loss, danger, inner conditions (depression, anxiety); (b) states with positive or neutral factors: such as abundance, security, moderation, inner conditions (optimism, certainty).

(4) *Qualifiers*: the specific aspects of the first three categories: namely, temporal characteristics (past, present, future, duration of an episode); contingency (degree of certainty); intensity (strength of items in story); negation (any type of denial); subordination (any means-end relationship); causality (any causal relationship).

Any word or statement in the subject's stories may, of course, be classified under one or more of the major categories and their subdivisions, as given above. This method of analysis is exceedingly detailed and laborious; but its author urges that it often yields insights not otherwise obtainable.

Comparison of the two foregoing analytic outlines, and comparison of these with other schemes, make it obvious that the TAT is not an objective test in the sense that tests of intelligence and specific aptitude are objective; first, because the details of the stories are largely evaluated and classified rather than scored; and second, because there are as yet no uniform standards or criteria which can be objectively applied for evaluating or scoring. We are not maintaining that an instrument like the TAT should or can conform to the standards and criteria of objectivity in scoring and reporting employed with other types of psychological tests. But the fact that the TAT does not so conform explains why its interpreters report their results in somewhat different terms and why its use requires acute psychological insights on the part of interpreters, particularly into the psychology of human needs and motivation. It is to be hoped, however, that before too long, after evaluations of the effectiveness of the several analytic schemes in

actual practice have been made, a single plan of analysis will be achieved.

Regardless of the particular scheme used in making an analysis, the results are interpreted as representing, literally or symbolically, tendencies and traits of the subject's personality, belonging to his past or present, or projected into the future. The results are interpreted, also, as representing, literally or symbolically, effective forces in the subject's environment, his views of the world, his past experiences, his anticipations of the future. The results and conclusions reached by any analysis are to be used as an hypothesis to be checked against other sources of information and as a starting point for further psychological interview, counseling, or treatment.

Reliability.⁴⁹ TAT reliability has been studied in three ways:

1. extent of agreement among interpreters of the same stories in regard to traits of the persons examined;
2. similarities between stories on repeated examinations of the same persons;
3. split-half method, correlating frequency and intensity of needs expressed in the stories.

Agreement among Interpreters. The first of these methods is dependent, in part, upon the reliability of the interpreters. Published reports indicate that agreement among interpreters is greater if they have had similar backgrounds in regard to training and if they use similar systems of analysis and scoring. This situation, of course, is only as it should be; namely, the interpreters approximate uniformity in their criteria of analysis and ratings. Such a situation is an approximation to the prescribed and uniform scoring standards and criteria provided for tests of intelligence and of specific aptitudes.

Studies of agreement among interpreters, using for the most part

⁴⁹ See, as examples: R. Clark, *A Method of Administering and Evaluating the Thematic Apperception Test in Group Situations*, Genetic Psychology Monograph, No. 3, 1944; A. W. Combs, "The Validity and Reliability of Interpretation from Autobiography and Thematic Apperception Test," *Journal of Clinical Psychology*, Vol. 2, 1946, pp. 240-247; R. Harrison and J. B. Rotter, "A Note on the Reliability of the Thematic Apperception Test," *Journal of Abnormal and Social Psychology*, Vol. 40, 1945, pp. 97-99; S. L. Garfield and L. D. Eron, "Interpreting Mood and Activity in TAT Stories," *Journal of Abnormal and Social Psychology*, Vol. 43, 1948, pp. 338-345; M. Mayman and B. Kutser, "Reliability in Analyzing TAT Stories," *ibid.*, Vol. 42, 1947, pp. 365-368.

rank-order correlation and the coefficient of contingency, have reported coefficients ranging from approximately $+ .30$ to $+ .90$.

When percentages of agreement among interpreters was the method used, clinicians have agreed completely on interpretations of from 50 to about 75 percent of the stories. In addition, there was *essential* (though not detailed agreement) in from 10 to 25 percent. On the remaining stories there was only partial agreement.

The foregoing data reflect, in part, differences in systems of analysis and differences in ability and experience of interpreters. The results of these researches are influenced also by the complexity and often intangibility of elements in the stories. Considering the nature of the problem, the reliability results obtained by this method are very encouraging.

Test-retest. Reliability data obtained by means of the test-retest method are affected by the stability of the personalities being examined and by personality changes as a function of time. The greater the time interval, the lower reliability we may expect, because there will be more opportunity for the influence of intervening forces. In view of the fact that it is to be expected that personalities often will change with the passage of time as developmental conditions change, especially in the case of children, of adolescents, and of persons who are clinical subjects, the more significant reliability data are those based upon degree of agreement of judges and those of test-retest after a brief interval. Tomkins⁵⁰ reports a reliability coefficient of $+ .80$ after an interval of two months, using fifteen young women as subjects; $+ .60$ after an interval of six months, using a different group of fifteen comparable subjects; and $+ .50$ after ten months for a third group. While these last two coefficients indicate that there were significant changes in ratings of some subjects, the indexes represent group trends, and do not signify that all subjects showed important changes. In fact, it was found that the time intervals between tests had little or no effect upon the ratings of stable personalities.

Split-half Method. Using the split-half method, Sanford⁵¹ reported reliability coefficients of $.48$ and $.46$. The responses were quantified by analyzing the stories for frequencies and rating intensities of "needs" and "press" elicited by the pictures. As reliability coefficients, these

⁵⁰ Tomkins, *op. cit.*, pp. 6-8.

⁵¹ R. N. Sanford et al., *Physique, Personality, and Scholarship*, Monograph of the Society for Research in Child Development, No. 8, 1943, p. 263.

are ordinarily too low to be of considerable significance. Under the circumstances, however, it is a matter of surprise that the coefficients are even this high, because the split-half method is inappropriate for the TAT. The method is inappropriate because not all of the pictures necessarily assess the same "needs" and "press," nor are they intended to do so. Although several pictures may to some extent elicit the same or similar "needs" and "press," each is intended to have a different stimulus value. The split-half method has been virtually abandoned in the study of TAT reliability.

Tomkins reported results obtained with one person who was studied intensively over a period of ten months, from whom about four hundred stories were obtained, using TAT pictures and others also. One interpreter rated the traits revealed in half the stories, while another rated the other half. The obtained correlation coefficient was $+.91$.

Of the several methods used in studying reliability of the TAT, interpreter agreement is the most useful and most significant; for in employing a projective test, what we want to learn primarily and as reliably as possible is the content and organization of a personality as it is *at the time of examination*. Whether that personality will remain the same and whether we can predict what that personality will be in time will depend upon the degree of stability of the individual being tested and upon the nature of the factors in his environment.

Validity. Essentially, several methods of studying TAT validity have been employed:

1. comparison with past histories and/or with results obtained through an intensive case study employing a variety of techniques;
2. comparison of characteristics of known individuals or groups with their TAT records;
3. comparison of TAT findings with other clinical materials: the subject's Rorschach record, dreams, or psychoanalytic interpretations;
4. experimentally produced changes.

Since most of the studies reported below do not fall exclusively in one category or another, they are not being presented according to the foregoing numerical classification.

Matching. Of these, the first method has been explored more than the others, Harrison's studies being regarded as among the most sig-

nificant.⁵² Forty patients at a mental hospital were given the TAT, without any prior knowledge, on the part of the examiner, regarding their histories. On the basis of their stories and behavior during the testing, Harrison drew his inferences concerning the personality development, traits, attitudes, level of intelligence, personal problems, and conflicts of each subject. These inferences were checked by another person against the hospital records. A correlation of $+ .78$ was obtained between estimated and obtained IQ's; 82 percent of the inferences were correct; in somewhat over 75 percent of the cases, diagnostic classification was correct, using the *major* categories; when 18 cases were classified into clinical *subgroups*, the percentage of agreement with clinical classification was 67. In order to eliminate inferences drawn from observation of the subject's behavior during the testing session, Harrison had another examiner give the TAT to 15 patients; then he made a "blind" analysis himself. In this instance, his inferences were 74 percent correct, when compared with already known biographical and personality data. This drop in correspondence (from 82 to 74 percent) indicates the value of using behavioral observations in conjunction with projective-test results.

In another and unusual matching experiment, an adaptation of the TAT was administered to groups of Navaho and Hopi Indians.⁵³ On the basis of their stories, "blind" interpretations were made of the personality traits of the people of these two cultures. Anthropologists familiar with the two Indian societies found the personality analyses based upon TAT results to be in essential agreement with their facts regarding these North American Indian cultures.

A preliminary investigation in India, using a modified and adapted form of the TAT, yielded highly promising results regarding the test's applicability to the study of social problems in that culture. The results make a positive contribution also to the question of validity; for a very high percentage of the stories dealt with the basic problems of survival (Murray's "succorance"), intra-family relationships (between the wife and the family of the husband in particular; Murray's "submis-

⁵² R. Harrison, "Studies in the Use and Validity of the Thematic Apperception Test with Mentally Disordered Patients. II. A Quantitative Validity Study," *Character and Personality*, Vol. 9, 1940, pp. 122-133; III. "Validation by Blind Analysis," *ibid.*, pp. 134-138. Also Harrison and Rotter, *op. cit.*, footnote 49.

⁵³ W. E. Henry, *The Thematic Apperception Technique in the Study of Culture-Personality Relations*, Genetic Psychology Monographs, Vol. 35, 1947, pp. 3-135.

sion"), and the need for education and skills to improve one's lot (Murray's "achievement" and "acquisition").⁵⁴

Known Groups. Satisfactory results were obtained when the stories of diagnosed groups, of known characteristics, were analyzed in detail to determine if significant differences existed among them. The results showed that such do exist between the following classifications, consisting of individuals who were relatively clear cases in each instance: conversion hysteria, anxiety hysteria, obsessive-compulsive neurosis, brain disease, and head-injury cases.⁵⁵

Rapaport, Gill, and Schafer made a qualitative analysis of TAT responses of clinically diagnosed individuals. They found trends in responses which are significant for diagnostic purposes with groups, such as the depressive, the paranoid, the schizophrenic.⁵⁶

TAT responses were found to have considerable validity in the personality studies of adolescent delinquents in a juvenile court.⁵⁷

The TAT has been used also, and with favorable results for the study and diagnosis of known non-clinical groups. Among these are the selection of potential leaders from among officer candidates in the armed forces,⁵⁸ and the differentiating TAT responses of prejudiced and non-prejudiced persons.⁵⁹

Comparisons with Autobiographies. Data obtained from autobi-

⁵⁴ These unpublished studies were made by one of the author's former graduate students, a native and life-long resident of India.

⁵⁵ E. R. Balken and J. H. Masserman, "The Language of Fantasy: III. The Language of the Fantasies of Patients with Conversion Hysteria, Anxiety State, and Obsessive-Compulsive Neurosis," *Journal of Psychology*, Vol. 10, 1940, pp. 75-86; H. Renaud, "Group Differences in Fantasies: Head Injuries, Psychoneurotics, and Brain Diseases," *Journal of Psychology*, Vol. 21, 1946, pp. 327-346; A. H. Davison, "A Comparison of the Fantasy Productions on the TAT of Sixty Hospitalized Psychoneurotic and Psychotic Patients," *Journal of Projective Techniques*, Vol. 17, 1953, pp. 20-33; C. H. Saxe, "A Quantitative Comparison of Psychodiagnostic Formulations from the TAT and Therapeutic Contacts," *Journal of Consulting Psychology*, Vol. 14, 1950, pp. 116-127.

⁵⁶ D. Rapaport, et al., *Diagnostic Psychological Testing*, Vol. II, Chicago: Year Book Publishers, 1946, pp. 439 ff.

⁵⁷ A. A. Hartman, *An Experimental Examination of the Thematic Apperception Technique in Clinical Diagnosis*, Psychological Monographs, No. 303, 1949.

⁵⁸ H. A. Murray and M. I. Stein, "Note on the Selection of Combat Officers," *Psychosomatic Medicine*, Vol. 5, 1943, pp. 386-391.

⁵⁹ E. Frenkel-Brunswick, "Dynamic and Cognitive Categorization of Qualitative Material: I. General Problems of the TAT," *Journal of Psychology*, Vol. 25, 1948, pp. 253-260.

ographies have been compared with TAT interpretations. Findings demonstrate that some of the pictures (about 30 percent) elicit stories that reflect past history better than do others and that, in this respect, the most useful pictures are those that contain characters with whom the subject is able to identify.⁶⁰

Comparisons with Dreams. Since psychoanalytic theory holds that dreams are a medium for expressing fears, wishes, etc., that are repressed and inhibited, and since the TAT provides an opportunity for expressing and elaborating these, it was to be expected that degree of correspondence between dream content and TAT stories would be investigated. Though not all the themes of the TAT responses appeared in dreams of the subjects thus studied, the very few reports available state that the extent of similarity was great enough to give added evidence of TAT validity.⁶¹ In this connection, the highly subjective character of interpreting dreams and dream symbolism must be kept in mind.

Agreement with Rorschach Findings. Several investigators have reported sufficiently close agreement between the Rorschach and the TAT as additional evidence of the validity of the latter.⁶² To use the inkblot test as a criterion of validity for another test is to attribute satisfactory validity to the Rorschach. This method must be evaluated in the light of the discussion of Rorschach reliability and validity.

Intensive Study of Individual Cases. Morgan and Murray report that the TAT stories of one patient indicated all the major characteristics revealed by five months of psychoanalysis.⁶³ This highly successful outcome is due in part to the fact that the psychologists making both the TAT and the therapeutic studies were psychoanalytically oriented. This fact does not minimize the findings; it emphasizes the possibilities of the TAT when uniform analytical systems are used with the test and with outside behavior.

Tomkins made an intensive study of one person, consisting of seventy-five hours of psychological interview and testing. He con-

⁶⁰ A. W. Combs, *op. cit.*

⁶¹ S. B. Sarason, "Dreams and Thematic Apperception Test Stories," *Journal of Abnormal and Social Psychology*, Vol. 39, 1944, pp. 486-492.

⁶² W. E. Henry, *op. cit.*; and R. Harrison, "The Thematic Apperception and Rorschach Methods of Personality Investigation in Clinical Practice," *Journal of Psychology*, Vol. 15, 1943, pp. 49-74.

⁶³ C. D. Morgan and H. A. Murray, *op. cit.*

cludes that the results of the intensive study disclosed no material inconsistent with his TAT stories and analysis.⁶⁴ On the whole, it was found that the TAT and other methods supplement one another, that each contributes something to an understanding of the personality not revealed by the others.

Experimental Changes. There have been few reports of experimentally produced changes in connection with the TAT. The principle of this method is that validity is shown to the degree that experimentally induced changes correspond with TAT changes. In one such experiment the need of "achievement" was selected for study with college students, with whom four pictures were used, administered under experimentally controlled conditions.⁶⁵ Different subjects performed under conditions called "relaxed," "neutral," "failure," and "success-failure." The "relaxed" state was created by telling the subjects that the test is merely experimental; the "neutral" by urging them to do their best, though the test is experimental; the "failure" by creating a sense of failure on previous paper and pencil tests; the "success-failure" by creating a sense of success and of failure on previous pencil and paper tests.

The "relaxed" state was interpreted as being least motivating and the "failure" state most motivating as regards the need for "achievement." When this hypothesis was tested, as related to scoring categories relevant to "achievement," significant differences under the four experimental conditions were found (at the 5-percent level or better) for many of the categories: e.g., increase in achievement reports, of deprivation related to achievement, acting to achieve a goal, projecting a goal.

It appears, then, that temporarily induced ego-involving tasks can influence TAT records. There is a basic distinction, however, between experimentally produced, transitory conditions and their effects upon responses, on the one hand, and actual, more or less permanent personality traits and content, on the other. The TAT is intended to assess the latter, primarily. The foregoing type of experiment, therefore, has only very limited significance as a method of estimating this instru-

⁶⁴ S. S. Tomkins, "Limits of Material Obtainable in the Single Case Study by Daily Administration of the Thematic Apperception Test," *Psychological Bulletin*, Vol. 39, 1942, p. 490.

⁶⁵ D. C. McClelland, et al., "The Projective Expression of Needs: IV. The Effect of the Need for Achievement on Thematic Apperception," *Journal of Experimental Psychology*, Vol. 39, 1949, pp. 242-255.

ment's validity. This type of experimental method can indicate only the test's degree of sensitivity to temporary and artificial conditions under which a person performs.⁶⁶

Evaluation of the TAT. The Thematic Apperception Test has sufficiently commended itself, from the point of view of psychological theory and practice, to become one of the most widely used of the projective techniques. It is being employed primarily in clinical studies of the maladjusted and the abnormal. At the same time, it is being used in the study of personality traits of selected normal groups (e.g., college students), of groups having particular attitudes (e.g., anti-labor), and of cultures different from our own. All these applications of the TAT have contributed to a better understanding of the personalities with which they have been concerned. In the main, the principal use and value of this instrument, at present, is to provide valuable material supplementary to other sources of information.

In responding to the thematic-apperception test-situation, the subject is free from social tensions that often accompany the early psychological interviews. Consequently, TAT responses provide effective starting places for interview and treatment; they are thus time savers and facilitators in the entire process.

The range of material in the test—permitting wide quantitative and qualitative individual differences in expressions of wishes, fantasies, frustrations, modes of adjustment—is one of its assets as a projective method. At the same time this wide range has thus far proved to be a weakness, in that procedures in administering, scoring, and interpreting vary, depending upon the conceptual system of the user and the purpose for which he is employing the test. Further use and analysis should determine the most desirable standard procedure and the most valid organization and interpretation of responses. The achievement of these ends demands, however, a more adequate systematic analysis of the psychological process underlying TAT reports. Improvements in analyzing and interpreting responses should contribute to greater validity of the test.

Validity can be improved, also, by more analyses of relationships between details of stories in conjunction with known behaviors and

⁶⁶ For a very useful summary see G. Lindzey, "TAT: Interpretive Assumptions and Related Empirical Evidence," *Psychological Bulletin*, Vol. 49, 1952, pp. 1-25.

traits of the subjects who tell them; that is, a close study of the *elements* within the stories that contribute to validity and of those that do not. Concurrently, it will be necessary to analyze and describe the kinds of *situations* (the organized "psychological fields") in which the TAT is more or less effective or ineffective. Beginnings have been made in these directions of research. If they prove fruitful, we should then be able to indicate not only validity in general but also validity in particular with respect to certain specified personality patterns or syndromes, and with respect to certain kinds of situations.

It will be necessary, furthermore, to analyze TAT reports—in fact, all projection-test responses—in the light of significant determinants of behavior and personality development of the subjects. Such determinants are age, sex, special occupational training, and social, economic, and cultural status and values (broadly, one's caste and class status). Upon the basis of such analyses it probably will be found that formulas and categories of scoring and interpreting will require modification and that the test itself will need revision and extension. Indeed, it is already becoming clear that separate thematic apperception tests are needed for some groups, such as children and adolescents. These are described in the next chapter.

PROJECTIVE METHODS: VARIOUS

IT WAS inevitable that a variety of projective devices should appear after the Rorschach and the TAT had found such wide favor and usefulness among psychologists. Although many of the newer tests are being utilized, more or less, and are proving useful to some degree in a limited number of situations, at present they must be regarded as tentative and in their early stages of development. In this chapter, some of these will be described and evaluated in order to give the reader a more comprehensive view and appreciation of this field of psychological testing. Of these, all but one (word association) appeared subsequent to the Rorschach and the TAT. In addition, we shall describe several other projective methods that are not tests in the usual sense but that have been employed for many years and are still widely used and are being improved (e.g., play, story telling, finger painting).

WORD ASSOCIATION TESTS

This projective method has a long history in psychological experimentation, dating from the work of Francis Galton published in 1879.¹ For many years after that, word association was experimentally studied in psychological laboratories. With the growing interest in psychoanalysis, after 1900, the word association method received increasing attention as a possible clinical technique. Jung and his colleagues, beginning about 1906, made extensive investigations of the

¹ "Psychometric Experiments," *Brain*, Vol. 2, 1879, pp. 149-162.

technique for use as a quick means of detecting "complexes."² Other clinicians, preceding and following Jung, were likewise concerned with word associations as a method of diagnosis.

Jung's list of one hundred words was selected as representing the common emotional "complexes." The subject is told that the examiner will speak a series of words, one at a time; after each word, the subject is to reply as quickly as possible with the first word that comes to his mind; there is no right response, no wrong response. The examiner records the reply to each stimulus word, the reaction time, and any unusual speech or behavior manifestations accompanying a given response. Replies to stimulus words that are emotionally toned for the subject generally have a longer reaction time and may also evoke physiological changes (e.g., in respiration, flushing, blood pressure), restless movements, coughing, laughing, and mild speech impediments. Jung proceeded on the principle that whenever a stimulus word was relevant to an emotional disturbance of the subject, an irregular response would result. The content of the responses, their reaction time, and attendant conditions were analyzed for the discovery of emotional tensions, inferred from the classes of words to which noteworthy responses were given. The inferences then serve as starting places for further psychological interview.

The best known of the word-association tests in the United States is the one devised by Kent and Rosanoff for the purpose of differentiating between the mentally ill and the normal.³ Unlike Jung, they used words that were not intended to indicate personal emotional problems but were, rather, neutral in character and were to provide diagnostic evidence on the basis of the proportion of common (normal) responses to the uncommon (abnormal). Recording reaction times and behavior observations were not part of their technique. Determination of normal and abnormal replies rested upon the frequencies of word associations of 1000 normal and about 250 psychotic subjects. The resulting table of response frequencies provided the percentages of most common, less common, and uncommon replies. The percentage of responses falling in each category, it was expected, might differ-

² C. G. Jung, *Studies in Word-Association*, trans. by M. D. Eder, London: Heinemann, 1918; also, "The Association Method," *American Journal of Psychology*, Vol. 21, 1910, pp. 219-269.

³ G. H. Kent and A. J. Rosanoff, "A Study of Association in Insanity," *American Journal of Insanity*, Vol. 67, 1910, pp. 37-96, 317-390.

entiate the normal from the abnormal. It was found, however, that the associations did not distinguish clearly enough between the two groups, although the results were at times useful as additional evidence in the case of a particular person.

The Kent-Rosanoff word list is given below.

- | | |
|---------------|--------------|
| 1. table | 41. high |
| 2. dark | 42. working |
| 3. music | 43. sour |
| 4. sickness | 44. earth |
| 5. man | 45. trouble |
| 6. deep | 46. soldier |
| 7. soft | 47. cabbage |
| 8. eating | 48. hard |
| 9. mountain | 49. eagle |
| 10. house | 50. stomach |
| 11. black | 51. stem |
| 12. mutton | 52. lamp |
| 13. comfort | 53. dream |
| 14. hand | 54. yellow |
| 15. short | 55. bread |
| 16. fruit | 56. justice |
| 17. butterfly | 57. boy |
| 18. smooth | 58. light |
| 19. command | 59. health |
| 20. chair | 60. Bible |
| 21. sweet | 61. memory |
| 22. whistle | 62. sheep |
| 23. woman | 63. bath |
| 24. cold | 64. cottage |
| 25. slow | 65. swift |
| 26. wish | 66. blue |
| 27. river | 67. hungry |
| 28. white | 68. priest |
| 29. beautiful | 69. ocean |
| 30. window | 70. head |
| 31. rough | 71. stove |
| 32. citizen | 72. long |
| 33. foot | 73. religion |
| 34. spider | 74. whiskey |
| 35. needle | 75. child |
| 36. red | 76. bitter |
| 37. sleep | 77. hammer |
| 38. anger | 78. thirsty |
| 39. carpet | 79. city |
| 40. girl | 80. square |

- | | |
|-------------|--------------|
| 81. butter | 91. moon |
| 82. doctor | 92. scissors |
| 83. loud | 93. quiet |
| 84. thief | 94. green |
| 85. lion | 95. salt |
| 86. joy | 96. street |
| 87. bed | 97. king |
| 88. heavy | 98. cheese |
| 89. tobacco | 99. blossom |
| 90. baby | 100. afraid |

A number of other word lists, having more or less in common with the foregoing, are available. The most recent of these, representing a revival of clinical interest in this technique, is that provided by Rapaport, Gill, and Schafer.⁴ They intend their list as an aid in clinical diagnosis, and for estimating the degree of maladjustment and impairment of thought organization. Their list is heavily loaded with stimulus-words presumably of analytical significance, especially in respect to psychosexual matters. While their statistical evidence is quite tentative, the work and interpretations of these investigators has the merit of being based upon an analysis of what they believe to be the psychological processes involved in a word association test. Their psychological rationale rests upon three aspects of associative response which is considered to be a thought process: (1) its memory aspect, chiefly in terms of the emotional factors that affect the process of associative recall; (2) concept formation, in terms of relevance of response to stimulus word; (3) "anticipation" aspect, in terms of the common or popular character of the conceptual responses, based presumably upon the ability of the subject to take a set or attitude from the examiner's instructions and from the character of the test words themselves, and to produce a sensible response.

A very recent modification of the word association technique is the *homographic* free association test. A homograph is a word spelled exactly like another but with a different meaning and a different derivation. An example is the word "base" which means "foundation" and is the homograph of "base" meaning "wicked." Thurstone's homographic test, for instance, uses a list of words, each of which has two frequent associations: namely, a social, interpersonal one and a literal,

⁴ *Op. cit.*, Vol. 2, Chapter 2.

physical one.⁵ The word "revolution" is one of such a list. The subject is asked to respond to each word with a synonym or a short phrase. A person's associative response to "revolution," for example, can have social or physical relevance: namely, "a political upheaval" or "the turn of a wheel." The purpose of such a restricted list of words would be to identify persons who are the more strongly oriented toward their social environment. This is shown by the number of responses that are essentially social in nature rather than physical or literal.

The factors influencing associations to word lists are many and must be taken into account in the construction and utilization of response-frequency tables, and in the interpretations of responses.⁶ These factors include not only the possibility of "complexes" and thought impairment, but differences in word usage resulting from regional, cultural, and socio-economic membership; from level of general intelligence; and from age level. Generally, therefore, the significance of any word responses must be sought in the experiences of the individual giving them.

In the study of personality, both as to content and structure, it is not likely that word association tests will have nearly the wide use and importance of the Rorschach and the thematic apperception tests.

PICTURE TESTS

The Travis-Johnston Projection Test.⁷ This 1949 adaptation of the thematic apperception method is intended specifically for the exploration of parent-child relationships. The test consists of two sets of drawings for use with children between the ages of 4 and 15 years—one set for girls and one for boys, each having 44 pictures. The pictures have been devised to evaluate the ways a child manages his "basic strivings in the face of cultural demands"; that is, his methods of becoming socialized and his degree of success. In selecting the situations to be portrayed, the authors assumed that ". . . certain wants and wishes from the earliest days of life will run counter to the socializing process set by the family constellation." This assumption, how-

⁵ L. L. Thurstone, *Word Associations with Homonyms*, University of Chicago Psychometric Laboratory, No. 79, 1952.

⁶ See, for example, D. Rapaport, *Emotions and Memory*, Baltimore: Williams and Wilkins, 1942; A. D. Tandler, "Significant Features of Disturbance in Free Association," *Journal of Psychology*, Vol. 20, 1945, pp. 65-89.

⁷ Distributed by Griffin-Patterson Co., 544 W. Colorado Blvd., Glendale, California.

ever, is not necessary for the use of the test or for the interpretation of a subject's reports.

The pictures are drawn in the form of sketches, omitting distinguishing facial features and unnecessary environmental detail; they are



FIG. 20.1. A thematic apperception test for the exploration of parent-child relationships. From the Travis-Johnston Projection Test, Griffin Patterson Co. (By permission.)

thus believed to be usable for the several cultural and socio-economic groups of our population. They portray adults and children of both sexes in numerous and varied situations and relationships which deal with important and potentially and actually troublesome areas in the socialization of children. The areas selected are sibling rivalry, child-parent rivalry, discipline, eating, sleeping, toilet training, cleanliness,

and sexual development. This test is distinguished by the fact that its situations are structured for age, sex, types of activity, combinations of characters and objects, and specification of areas to be explored.

Being quite new, this instrument is in process of development; and, as its authors state, more time is needed for its satisfactory evolution. This test does have promise of value in supplementing children's case histories, disclosing dynamics of behavior, identification of behavior areas that are particularly troublesome, and indicating the need for and success of therapy.

Thompson modification of the TAT.⁸ Clinical experience with the Murray TAT convinced some psychologists that Negroes very often were unable to identify themselves with the stimulus figures of white persons. These subjects gave reports (stories) which were largely simple, matter-of-fact descriptions of the persons and objects in the pictures. They did not show empathy based upon the interpersonal relationships being portrayed. Yet the original TAT itself is based upon the principle that greatest identification will occur when there is the greatest number of symbolic elements in the picture relevant or common to the subject being examined. Hence, if a picture does not reflect the experience and culture of the examinee, there will be little or no identification and the test will yield only simple descriptions or will become a mere verbal exercise.⁹

Children's Apperception Test.¹⁰ This test, for children of 3 to 10, consists of ten pictures in which the characters are animals rather than human beings. The pictures show these animals in commonplace *human* situations and relationships: eating, sleeping, shopping, being punished, etc. The assumption underlying the use of pictures of animals is that children will more readily identify with them than with persons. In support of this, the authors cite only Sigmund Freud's widely known report on "The Phobia of a Five Year Old."

The authors state in their manual: "The pictures were designed to elicit responses to feeding problems specifically, and oral problems

⁸ Prepared by C. E. Thompson. Published by Harvard University Press, 1949.

⁹ For a criticism of the underlying assumption of this revision see B. F. Riess et al., "Further Critical Evaluation of the Negro Version of the TAT," *Journal of Projective Techniques*, Vol. 15, 1951, pp. 394-400.

¹⁰ By L. Bellak and S. S. Bellak. Published by C.P.S. Co., P.O. Box 42, Gracie Station, New York 28, N. Y. Second edition, 1950.

generally; to investigate problems of sibling rivalry; to illuminate the attitude toward parental figures and the way in which these figures are apperceived; to learn about the child's relationship to the parents as a couple. . . ." The pictures are intended to elicit, also, the child's fantasies regarding aggression and the adult world, and his methods of responding to and dealing with his problems of growth.

The themes of the pictures are derived primarily from problems and relationships suggested by psychoanalytic principles of development and behavior; and interpretations of the stories rest upon symbolic significance attributed to the content. Although psychoanalytic principles guided the authors of this test in its construction, it is possible to use other systems in the interpretation of responses.

While reports from clinicians indicate that this test can be helpful in some instances, when used by a skillful person, acceptance of the basic assumptions must await actual demonstration through research. A number of questions are yet to be answered: Do children in general identify more readily with animals? ¹¹ What are the relationships between responses and (1) age, and (2) socio-economic status? How are responses influenced by intelligence level? ¹² Are there significant sex differences?

Symonds Picture-Study Test.¹³ This set of pictures is designed for use with adolescent boys and girls. The cards depict a large number of situations and interpersonal relationships in which individuals of this developmental stage commonly find themselves. The stories told in response to the pictures are to be analyzed for the psychological forces indicated by them. The forces are among those commonly described in dynamic psychology:

¹¹ For negative results, see K. R. Biersdorf and F. L. Marcuse, "Responses of Children to Human and to Animal Pictures," *Journal of Projective Techniques*, Vol. 17, 1953, pp. 455-459.

¹² Responding to pictures is one of the test items long used in the Binet scales. At age 3-6, only "enumeration" is expected. The next stages are "description" and "interpretation" (no longer included in the S-B). Since this is the course of development, it seems that "identification" may not be expected ordinarily at the first two levels, at any rate. This position was supported, at least tentatively: N. A. Kaake, "The Relationship between Intelligence Level and Responses to the Children's Apperception Test," Unpublished M.A. Thesis, Cornell University, 1951.

¹³ By P. M. Symonds. Bureau of Publications, Teachers College, Columbia University, 1948.

hostility and aggression	ambition and striving for success
love and erotism	conflicts
ambivalence	guilt
punishment	guilt reduction
anxiety	depression, discouragement, despair
defenses against anxiety	happiness
moral standards and conflicts	sublimation

Tentative norms indicate that the five themes most frequently obtained deal with family relationships, aggression, economic concern, punishment, and separation. Informal reports thus far available on this instrument suggest that it and similar specialized tests have possibilities of considerable usefulness; but as is the case with most of the new devices, significant research findings have not yet been published.¹⁴

The Blacky Pictures.¹⁵ This set of eleven pictures, intended for ages 5 and over, is offered as a "measure of psychosexual development." Originally devised to test several Freudian hypotheses, the pictures are now used to learn the degree to which the subject has developed the various psychosexual traits. The pictures portray a young dog, Blacky, in situations intended to represent relevant experiences with three other dogs: namely, his father, his mother, and his sibling.

As in the case of the Children's Apperception Test, the Blacky pictures often anthropomorphize the animals by presenting them in obviously and characteristically human situations. The same questions that were raised regarding the former device may be asked of the latter.

Validation studies thus far published suggest "that there is something there . . . but do not necessarily indicate what it is or where it is."¹⁶ Obviously, then, much more research is needed before the nature and significance of findings can be determined.¹⁷

¹⁴ Although the Symonds pictures were drawn by a professional artist, they seem to portray characters who predispose the subject to respond with unhappy or "problem" stories. This writer's graduate students have characterized the pictured individuals as "sad," "sour," "unhappy," "starved," "gaunt."

¹⁵ By G. S. Blum. The Psychological Corporation, 1950.

¹⁶ G. Blum and H. Hunt, "The Validity of the Blacky Pictures," *Psychological Bulletin*, Vol. 49, 1952, pp. 238-250.

¹⁷ See also, A. Ellis, "The Blacky Test Used with a Psychoanalytic Patient," *Journal of Clinical Psychology*, Vol. 9, 1953, pp. 167-172; M. L. Aronson, "A Study of the Freudian Theory of Paranoia by Means of the Blacky Pictures," *Journal of Projective Techniques*, Vol. 17, 1953, pp. 3-19.

The Four-Picture Test.¹⁸ Although this test was completed in 1930, it was not made available until 1948. Eighteen years were devoted to experimental and clinical work before van Lennep felt it was ready for distribution. As a result, the manual is carefully prepared, complete, and explicit.

The test consists of four vaguely drawn colored pictures (reproduced from water colors) that are placed before the subject who is asked to write (or tell) *one story that arranges and combines all four pictures* according to his own choice. Each picture portrays a different situation: (1) being together with one other person (a room showing two persons, with a table between them. One person is standing and apparently gesturing); (2) being personally alone (a bedroom with a window behind the bed, and a very dim suggestion of a person in the bed); (3) being socially alone (a figure leaning against a lamp post on a rainy night); (4) being together with many others in a group (a tennis match, showing two players in the background and, in the foreground, two couples talking).

While only four pictures are used (in marked contrast to other picture apperception tests), the fact that they have to be combined into a single story provides opportunities for a great variety of responses. The uniqueness of this picture test, then, lies in the use of color, the small number of pictures, and the requirement of a single integrated story.

The van Lennep test is not quantitatively scored in any respect. Its interpretation is entirely qualitative, being concerned with the subject's "general attitude toward life." This attitude is given in descriptive terms and in terms of dynamic and systematic psychological principles of behavior and development.

In spite of the fact that the test manual presents no statistical data, van Lennep's instrument has been favorably received—at times enthusiastically—by psychologists who have used it. This reception is attributable to the careful and thorough psychological analysis of responses obtained during the test's long period of development.

¹⁸ By D. J. van Lennep. Martinus Nijhoff, The Hague, Netherlands, 1948. See also by van Lennep, "The Four Picture Test," Chapter 6 in *Introduction to Projective Techniques*, H. H. and G. L. Anderson (editors), New York, Prentice-Hall, 1951; E. S. Schneidman, "Some Comparisons Among the Four Picture Test, TAT, and Make a Picture Story Test," *Rorschach Research Exchange and Journal of Projective Techniques*, Vol. 13, 1949, pp. 150-154.

Make A Picture Story Test.¹⁹ This device, for use with adolescents and adults, combines the essential feature of the TAT (telling stories) with an innovation: namely, giving the subject an opportunity to construct his own pictured situation within the limits of materials provided by the test. These materials consist of 22 "background pictures" (achromatic, on cardboard); some ambiguous, others semi-structured, still others definitely structured. Among the backgrounds are, for example:

living room	street	forest
bedroom	camp	cave
bathroom	landscape	schoolroom

There are also 67 cut-out figures (65 of which are human and 2 animal) representing:

male adult	indeterminate as to sex	children
female adult	legendary and fictitious	a dog
minority groups	silhouette and blank faces	a snake

These figures are portrayed in a variety of postures and states.

The examiner selects one background picture at a time, asks the subject to place one or more figures of his own choosing against the background as they might appear in real life and then to tell a story about the scene he has created. Any of the figures may be realistically placed against any background; thus the principle here is that the individual may project whatever actions and relationships he wishes upon whatever persons he chooses.

After each story, the examiner conducts an inquiry, as in the case of the TAT. The complete responses are recorded and analyzed according to a rather detailed and elaborate scheme. Essentially, the stories are analyzed primarily for form and secondarily for content. The second of these—content—may be of almost unlimited variety, intended to elicit through these "fantasy productions" the subject's social adjustments and relationships. By analysis of form, the test's author means ". . . which figures are chosen, how many are chosen, where they are placed on the background, how they are handled by the subject, and what relationships they bear to each other." Form is analyzed and characterized in terms of forces operating on the subject

¹⁹ By E. S. Schneidman. Psychological Corporation, 1948.

(goals, drives, conflicts, values, ego ideals, etc.) and of modes of behavior (hostility, sexuality, aspiration, autonomy, etc.).

It is thought by some psychologists that the stories evoked by means of this test may be more spontaneous and richer in fantasy production than the TAT because the pictured situation has been created by the subject himself. On the other hand, it is possible that the subject will evade certain types of situations and problems, significant in his case, with which he would have to deal in one way or another in the TAT type of test.

The Michigan Picture Test.²⁰ This very recent instrument is the most systematically and thoroughly constructed thematic apperception test that has appeared since Murray's TAT. It is intended for use with children from eight to fourteen years of age. The test consists of sixteen pictures, only twelve being used with each child, the selection depending upon the subject's sex. The basic principle is the same as that of the original TAT: revealing needs through responses to pictures. Construction of the Michigan test was motivated by reports from many child guidance clinics that the TAT was not entirely suitable for children under fourteen years of age and that a special test, suitable for younger subjects, was necessary.

The authors state in their manual that the over-all purpose of their project was to ". . . investigate and measure the emotional reactions to children in the preadolescent and adolescent stages of development. . . . It was believed the test should be non-traumatic and yet tap the common conflict situations for this age group." To do this, they followed the methods of population sampling, behavior sampling, and response analysis prescribed for and very frequently found in the construction of tests of intelligence and specific aptitudes.

Results of the construction process indicated that certain test variables did effectively discriminate between groups of well-adjusted and groups of poorly-adjusted children. These variables constitute: (1) the "tension index," (2) "verb tense," (3) "direction of forces." The tension index is based upon the following four types of needs: ²¹

love: verbal expression, positive or negative, indicating affection, affiliation, attachment, friendship, or admiration;

²⁰ By G. Andrew, S. W. Hartwell, M. L. Hutt, and R. E. Walton. Science Research Associates, 1953.

²¹ *Manual*, pp. 66 ff.

extrapunitive: verbal expression of aggression toward an external object;

submission: verbal expression of defeat, resignation, passivity, compliance, obedience, acceptance of suffering without opposition;

personal adequacy: verbal reference, positive or negative, of happiness, strength, competence, or any reference to the temperament or physical characteristics of the human or animal figures in the story.

"Verbal tense" of responses is scored for frequencies of past, present, and future tenses. The hypothesis, quite empirical in origin, is that disproportionate emphasis upon each of the several tenses may indicate behavior as follows:

past tense: avoidance of conflict
 a regressive trend
 schizoid character structure
 submissiveness or isolation

present tense: compulsivity or pedantry
 personality disturbance
 effective intellectual functioning

future tense: anxiety
 disturbed but relatively mature personality
 inefficient intellectual functioning

"Direction of forces" refers to the action expressed in the story:

centrifugal direction (outward action)
centripetal action (inward action)
neutral (no direction indicated)

Well-adjusted children express both centrifugal and centripetal directions much more frequently than the poorly-adjusted.

A second group of variables, for which the stories are also scored, consists of those ". . . for which trends were indicated, but in which differences were not statistically significant." This group of variables includes the following:

psychosexual level: a measure of psychosexual maturity in orthodox Freudian terms;

interpersonal relationships: range and frequency of expressed interpersonal relationships;

personal pronouns: frequencies of the three personal pronouns as a measure of self-reference;

popular objects: most commonly referred to objects and persons;

level of interpretation: degrees of interpretation, from "no response," through enumeration and description to complex interpretation and inference.

The results obtained for these four variables were inconclusive as between well-adjusted and poorly adjusted groups; yet they were suggestive enough to warrant further investigation as possible differentiating criteria.

The superiority of this Michigan test is attributable to the fact that the pictures were designed, in the first instance, in accordance with well-defined concepts of psychological development and of the most significant environmental situations affecting development. As already stated, careful sampling procedures were used. Also, unlike practically all other recent apperception tests, the Michigan test presents extensive data regarding significance of group differences, norms for each of the scoring variables, and reliabilities.²² Of the instruments appearing since the original TAT, this test, therefore, appears to rest on the soundest base and to be the most promising for the study and assessment of children's personalities within the specified age groups.

Rosenzweig Picture-Frustration Study.²³ The full name of this test is "Picture-Association Study for Assessing Reactions to Frustration," although it is generally referred to by the first title given above. Consisting of twenty-four cartoon-like pictures, the test is intended to serve as a projective method to reveal the subject's characteristic patterns of response to common stress-producing situations that are regarded as important in normal and abnormal adjustment. Two forms are avail-

²² Reliability correlations of the four variables constituting the "tension index" scores obtained by two judges, grade 3: .67, .93, .93, 1.00; grade 5: .91, .97, .97, .98; grades 7 and 9: .70, .81, .98, .98. Fifteen cases in each group. Reliability correlations for "direction of forces" as scored by two judges, grade 3: .95; grade 5: .87; grades 7 and 9: .91. Ten cases in each group.

²³ By S. Rosenzweig. The Author, 8029 Washington St., St. Louis 14, Mo., 1944-1948.

able, one for children of ages from 4 to 14 years, and one for persons older than fourteen.²⁴

Each picture shows two persons involved in a mildly frustrating situation of common occurrence. The person at the left in each picture is represented as making a statement which either helps describe the frustration of the second individual or is itself actually frustrating the latter. The caption box above the person on the right of each picture is blank. Purposely omitted are the facial features and expressions in all pictures.

The task of the subject is to examine each picture and to write in the blank box the first appropriate response that occurs to him. (Young children, of course, will give oral responses to be written by the examiner.) The assumption of this test is that the subject identifies himself, consciously or unconsciously, with the frustrated individual in each situation and that his replies thereto are projections of his own ways of acting.

The situations presented are classified as of two kinds: ego-blocking and superego-blocking. The first are those "in which some obstacle, personal or impersonal, interrupts, disappoints, deprives, or otherwise directly frustrates the subject." The second type of blocking "represents some accusation, charge, or incrimination of the subject by someone else." As with other projective methods, an inquiry follows the recording of responses for purposes of clarification and possible elaboration.

Scoring of responses is based upon (1) direction of aggression and (2) reaction type. Under the first of these, three forms of expression are distinguished: (a) extrapunitive in which aggression is turned onto the environment; (b) introjective, whereby the subject turns the aggression upon himself; and (c) impulsive, in which aggression is evaded in an effort to gloss over the frustration. Type of

²⁴ S. Rosenzweig, "The Picture-Association Method and Its Application in a Study of Reactions to Frustration," *Journal of Personality*, Vol. 14, 1945, pp. 3-23; S. Rosenzweig, E. E. Fleming, and H. J. Clarke, "Revised Scoring Manual for the Rosenzweig Picture-Frustration Study," *Journal of Psychology*, Vol. 24, 1947, pp. 165-208; H. J. Clarke, E. E. Fleming, and S. Rosenzweig, "The Reliability of the Scoring of the Rosenzweig Picture-Frustration Study," *Journal of Clinical Psychology*, Vol. 3, 1947, pp. 364-370; S. Rosenzweig, "Revised Norms for the Adult Form of the Rosenzweig P-F-S," *Journal of Personality*, Vol. 18, 1950, pp. 344-346.

reaction also has three classes: (a) obstacle dominance, in which the barrier occasioning the frustration stands out in the response; (b) ego defense, in which the ego of the respondent predominates; (c) need persistence, in which the subject emphasizes the solution of the frustrating problem.

The responses to the frustration pictures are intended to show the individual's "frustration tolerance," which signifies the relative absence of observable disorganization in response to frustrating situations; or, adequacy and efficiency of response despite frustrations. Frustrations are common experiences; modes of adjustment to them are regarded as significant in understanding behavior and personality organization, in that these modes reveal one's techniques for coping with tensions. The Rosenzweig pictures are intended to provide a method of evaluating an individual's tendency to blame the source of his frustration (extrapunitive), or himself (intropunitive), or to treat the situation impersonally (impunitive).

The Picture-Frustration Study has found more than a little favor with clinicians; and, with modifications, it has been used also in psychological studies of groups and group attitudes.²⁵

Two questions, however, have been raised regarding the interpretation to be given the responses to this test. First, are the responses that are given to the pictured incidents, which represent rather mild frustrating situations, to be interpreted as indicative of the subject's typical responses to frustrating situations in actual life that are of basic significance to his personality? Second, Symonds has asked whether the subject's responses indicate what he actually *would* do, or what he thinks he *should* do, or *feels* he would *like* to do, but would not actually do.²⁶ Although these questions have not been answered definitely, the results obtained with this test indicate either: (1) how the subject probably would act in these kinds of situations, or (2) what he knows or thinks would be expected of him in such situations. Using

²⁵ For example, J. L. McCary, "Reactions to Frustration by Some Cultural and Racial Groups," *Personality*, Vol. 1, 1951, pp. 84-102; J. D. Holzberg and R. Posner, "The Relationship of Extrapunitiveness on the Rosenzweig P-F-S to Aggression in Overt Behavior and Fantasy," *American Journal of Orthopsychiatry*, Vol. 21, 1951, pp. 767-779.

²⁶ O. K. Buros, *Fourth Mental Measurement Yearbook*, 1953, p. 243. Also, M. M. Schwartz and L. Karlin, "A New Technique for Studying the Meaning of Performance on the Rosenzweig P-F-S," *Journal of Consulting Psychology*, Vol. 18, 1954, pp. 131-134.

the responses, then, as a basis of inquiry and interview, discovery of the person's level of frustration tolerance is facilitated.

The Szondi Test.²⁷ Devised by the psychiatrist Lipot Szondi, this test consists of six sets of facial pictures of mental patients, eight in each set. Each picture is on a separate card. Each set includes one photograph said to portray the face of each of the following: a homosexual, a sadistic murderer, an epileptic, an hysteric, a catatonic, a paranoiac, a depressive, and a manic. Deri recommends that the test be administered at least six times to the same subject. From each set, the subject is asked to select the two photographs he most likes and the two he most dislikes. After making his selection, the subject is asked to tell a story about each of the four most liked and the four most disliked, of the entire forty-eight pictures, in order to reveal specific ways in which they affect him. The subject, of course, is not given any clues as to the presumed type represented by each picture. The number of choices made of each category is said to be indicative of some of the subject's personality traits. Selection of homosexuals is interpreted as indicating the characteristics of the subject's psychosexual development; selection of sadistic murderers indicates aggressiveness; epileptics, a violent quality of behavior and thinking; hysterics, the emotional aspect of life; catatonics, the quality of narcissism and withdrawal; paranoiacs, expansive forces and creative abilities; selection of manics and depressives, the person's "mood-life."

The number and classification of choices (like or dislike) in each category, it is claimed, indicate strength of tendencies that are *latent* in the person. If few or no choices are made in a category, the interpretation is that the tendencies are already *overt* in the individual.

According to the Szondi interpretation, therefore, every person has all the tendencies specified by him! If certain of the pictures are rejected altogether, the traits they are presumed to represent are said to be overt. Selected but *disliked* pictures mean, we are told, that the subject is repressing or sublimating the tendencies which it is claimed they represent. Selected and *liked* pictures, we are told, represent

²⁷ S. K. Deri, "Description of the Szondi Test: a Projective Technique for Psychological Diagnosis," *American Psychologist*, Vol. 1, 1946, p. 239; also, *Introduction to the Szondi Test*, New York: Grune and Stratton, 1949. D. Rapaport, "The Szondi Test," *Bulletin of the Menninger Clinic*, Vol. 5, 1941, pp. 33-39.

tendencies that are acceptable to the subject's "super-ego" and that are available for expression. For the subject who submits to the Szondi Test, there seems to be no escape from having attributed to himself some degree of each of the eight traits.

This test is based upon two extremely unsound assumptions. The first is that photographs can in themselves represent and differentiate the several kinds of personalities named. This assumption sounds too much like the long-since rejected pseudo-science of physiognomy. The second premise is that a person's overt behavior derives from his inherited genetic elements (genes) and that his selections are "instinctively" determined and represent latent forms of behavior which seek expression. This second assumption is contrary to the widely held views of the roles of heredity and experience in personality development. It is also contrary to the modern principles of behavior, some of which have radically modified the concept of instinctive behavior while others have rejected it entirely.

This test is briefly described here as an example of an instrument that creates personality traits by fiat of its author. Thus far, a fairly large number of studies on the Szondi have been reported, their findings and conclusions being almost uniformly negative or their conclusions very skeptical.²⁸

VERBAL COMPLETION TESTS

Sentence-Completion Tests.²⁹ As the name indicates, these are tests whereby the individual is presented with a series of incomplete sentences, generally open at the end, to be completed by him in one or more words. They are somewhat like the word-association test in that the word or words used to complete the sentence follow from and are associated with the given part of the sentence. The sentence-completion test, however, is regarded as superior to word-association because the subject may respond with more than one

²⁸ Buros, *op. cit.*, pp. 255 ff.; H. P. David et al., "Qualitative and Quantitative Szondi Diagnosis," *Journal of Projective Techniques*, Vol. 17, 1953, pp. 75-78; M. Fleischmann, "The Discriminative Power of the Szondi Syndromes," *Journal of Consulting Psychology*, Vol. 18, 1954, pp. 89-95.

²⁹ See M. L. Hutt, "The Use of Projective Methods of Personality Measurement in Army Medical Installations," *Journal of Clinical Psychology*, Vol. 1, 1945, pp. 134-140; A. R. Rhode, "Explorations in Personality by the Sentence Completion Method," *Journal of Applied Psychology*, Vol. 30, 1946, pp. 169-181; J. B. Rotter and B. Willerman, "The Incomplete Sentences Test," *Journal of Consulting Psychology*, Vol. 11, 1947, pp. 43-48.

word, a greater flexibility and variety of response is possible, and more areas of personality and experience may be tapped.

The content of a particular test and the nature of the sentences will depend upon the group of persons and the purposes for which they are intended. One test may be devised to learn about satisfactions and annoyances, likes and dislikes, fears and attractions. Another may be designed to reveal motives, needs, and environmental forces operating upon the subject. Still another may be concerned primarily or solely with feelings about one's situation, as home, community, friends, occupation, school. Others may be intended to discover certain psychological mechanisms, such as feelings of rejection, evidence of rationalization, methods of evasion. In other words, each sentence-completion test should be adapted to the particular situation in which it is to be applied.

Several sample items are the following:

- I worry over
- I feel proud when
- Other people usually
- I prefer to
- My father used to
- My hope is
- When I was a child

This projective method, at present in its early experimental stages, seem to have possibilities of use in the study of normal personalities as well as with clinical cases.

In this category, the two most fully described and most frequently used instruments are the *Sentence Completions Test*³⁰ (for ages 12 and over) and *The Rotter Incomplete Sentences Blank* (college form).³¹ Both are designed to estimate the subject's degree and areas of maladjustment, if any. Although both provide schemes for scoring responses, it appears that the greater usefulness of this type of test is for the indications of areas of maladjustment it may provide and for diagnostic clues. For instance, the responses of a ten-year-old boy to

³⁰ By A. R. Rhode and G. Hildreth. The Psychological Corporation, 1940-1947.

³¹ By J. B. Rotter. The Psychological Corporation, 1950. Manual in collaboration with J. E. Rafferty. Also, J. B. Rotter et al., "The Validity of the Rotter Incomplete Sentences Blank," *Journal of Consulting Psychology*, Vol. 18, 1954, pp. 105-111.

one set of completions were all neutral and indicative of satisfactory adjustment, except with regard to his school experiences.

On the whole, it appears that sentence completion tests evoke personality materials that are closer to the level of consciousness than those evoked by the Rorschach and the thematic apperception type.³² The sentence completions do, nevertheless, provide a basis for subsequent interview and counseling.

Projective Questionnaire. By means of this technique, the subject is given a series of questions to answer in his own way. Unless he is aware of the psychological significance of the questionnaire, he does not grasp the implications of the questions or of his answers to them. It is possible, by this method, to ascertain some information regarding the subject's emotional life, his values, his attitudes, and sentiments. Each question is intended to show which dynamisms might be operating in an individual's behavior; for example, suppression, projection, identification, and others. The specific questions to be included in any particular device will depend upon its purpose and upon the persons for whom it is intended. In any event, the value of this type of test, at present, does not lie in a numerical score; indeed, rating schemes have not yet been devised, since this technique is quite recent and has not been used extensively. The value of the projective questionnaire lies, rather, in the fact that the answers are interpreted as revealing certain traits and serve as a basis for psychological interview.

The following are items which were included in the questionnaire devised by the psychological staff of the Office of Strategic Services during World War II, to be used, among other tests, in the screening and job assignment of personnel. (See Chapter 21.)

It seems that no matter how careful we are, we all sometimes have embarrassing moments. What experiences make you feel like sinking through the floor?

What kinds of things do you most dislike to see people do?

What things or situations are you most afraid of?

If you were (are) a parent, what things would you try to guard your children against most carefully?

³² For example, E. Hanfmann, "Studies of the Sentence Completion Test," *Journal of Projective Techniques*, Vol. 17, 1953, pp. 280-294; J. Byers, "The Relationships between Sub-Cultural Group Membership and Projective Test Responses," Unpublished Ph.D. Thesis, Cornell University, 1954.

Story Telling and Story Completion.³³ Although this method has been used informally for some years,³⁴ the published work on it has not been very extensive or definitive. Several approaches have been employed, the most common with children being the following:

Retelling of children's popular stories, such as "The Three Bears."

Retelling the story the child likes best of all those ever heard or read.

Stories made up on specified themes: such as on a boy or girl, a father, a mother.

A story especially constructed for the purpose at hand, told to the children by a teacher; reproduced in writing for the teacher; later retold to the therapist, this being the "emotional version." (In evaluating such retold stories, the therapist must take into account the common phenomena of memory distortions.)³⁵

Users of these methods report that they are helpful in revealing a child's conflicts, aggressions, anxieties, wish fulfillment, affectivity, etc.

Among other methods which have been tried are these:

A situation is set up creating a moral conflict, the effects of which on the child are evaluated through the story he makes up about it and his methods of dealing with the conflict.

The child is asked to invent stories about favorite comic-strip characters; and these stories are regarded as indications of his "retreat into fantasy," on the one hand, or his realistic attitude toward his environment and interpersonal relations, on the other. The child is also asked to invent stories about disliked comic-strip characters as a means of discerning his personality tensions and for subsequent use in therapy, since these stories give evidence of emotional transference and are a means of communicating affectivity.

Story telling as a projective method has been tried also with adults.

Ordinarily the subject is asked to develop a story on a theme dealing with some aspect, or aspects, of his environment in order to bring out some of his strivings and modes of adjustment.

³³ E.g., H. Sargent, *An Experimental Application of Projective Principles to a Paper and Pencil Personality Test*, Psychological Monograph, Vol. 57, No. 5, 1944; B. A. Wright, "An Experimentally Created Conflict Expressed by Means of a Projective Technique," *Journal of Social Psychology*, Vol. 21, 1945, pp. 229-245.

³⁴ For example, P. Blos, *The Adolescent Personality*, New York: D. Appleton-Century, 1941.

³⁵ See F. C. Bartlett, *Remembering*, Cambridge: Cambridge University Press, 1932.

A little work has been done by combining music and story telling. Recordings of selected musical classics are played; the subject is told to report the images and themes he associates with these. The expectation is that the respondent's reports will indicate certain psychological states, such as fear, struggle, romance, animation, reverence, and others.

Story-completion techniques are of several kinds.

Dramatic situations are presented briefly and concisely; the subject is required to develop each into a skeleton of a short story.

The outline of a story is provided, it being the subject's task to write a narrative based upon it.

A brief incomplete story is presented, which the subject is to complete.

The individual is given a group of brief statements each of which presents a situation of emotional conflict, followed by two questions on how the central figure in the situation responded to it: "How did he feel?" "What did he do and why?"

A very recent elaboration and systematic development of this technique of story completion has been offered by H. D. Sargent, called *The Insight Test*. It ". . . is composed of a series of items . . . in which the bare outlines of a problem situation are stated and to which the subject is asked to respond by telling what the leading character did, and why, and how he felt about it. The nature of the material and the task itself, which is presented as a test of insight or ability to 'see into' the motives, actions, and feelings of others is designed to conform to two basic principles of a projective test: ambiguity of stimulus to which the subject must respond in his own personal terms, and direction of attention away from concern with self toward the task of reacting to something external."³⁶

This story completion test has the merit of having been provided with a system of analysis and scoring that is intended to yield interpretations based upon a set of common principles. The scoring system ". . . involves primarily a differentiation between expressions of feeling in response to the [items], and other types of expression which are regarded as serving the purpose of control, defense, and delay of unmodulated emotion. Phrases . . . are categorized both with reference to the fate of the aroused affect and to the mood or content ex-

³⁶ *The Insight Test*, New York: Grune and Stratton, 1953, p. 17.

pressed. Several indications of disturbance in thought or feeling are also recorded. Finally, certain relationships between expressions supposed to represent affect and defense, and within the quantities of different kinds of expression standing for different sorts of discharge and control, are computed."³⁷

It is obvious that as a testing technique, story telling and story completion present extreme difficulties in evaluation and interpretation; for, with the exception of Sargent's test, the situations are almost completely unstructured and fluid, and the possibilities of categorical analysis are almost unlimited. At present, and perhaps permanently, these techniques must be regarded as highly subjective in character and of use only for exploratory purposes when viewed against a background of other information on the individual's experiences, traits, and behavior.

DRAWING AND PAINTING³⁸

If volume of publication is a criterion, then it may be said that drawing and painting (including finger painting) are of increasing interest and significance as projective methods in the study of personality. Studies in the psychology of art as related to chronological age, as evidence of intelligence and specific aptitudes, as culturally influenced, and as aesthetic judgment, have provided the basis upon which investigations into the relationships between art and personality have been based. The point is that not only must the products of atypical groups be analyzed for elements and degree of communality, but the usual, normative products of drawing and painting of normal persons, as influenced by various factors, must be determined as background against which to project the creations of persons whose personalities are being studied for diagnostic or descriptive purposes. For example, in the study of artistic products of psychotics it is reported that, among other things, they are markedly similar to the drawings of children and some primitives, pronounced stereotypes and mannerisms are typical, representation is extremely non-realistic, and their meanings may be discovered in a manner similar to analysis of dream con-

³⁷ *Ibid.*, p. 26.

³⁸ E.g., P. Elksch, *Children's Drawings in a Projective Technique*, Psychological Monographs, Vol. 58, No. 1, 1945; W. Wolff, *The Expression of Personality: Experimental Depth Psychology*, New York: Harper, 1943; L. B. Murphy, "Art Technique in Studying Child Personality," *Rorschach Research Exchange and Journal of Projective Techniques*, Vol. 13, 1949, pp. 320-324.

tent. It is reported, also, that color and movement have diagnostic significance, as do size, line, form, and symbols used.

Finger painting has much in common with drawing and other forms of painting in personality study; but, in addition, its exponents emphasize the considerable diagnostic value of observing and analyzing the subject's approach to the situation and his behavior throughout.³⁹

Methods of obtaining productions vary among investigators and users of these techniques, as do their evaluations and interpretations, in some of which considerable free play is given by the interpreter to his particular theoretical biases, unsupported by scientific data. The great needs, before painting and drawing can come to occupy a reliable position in the area of projective methods, are that their psychological rationale shall be determined, their underlying psychological processes shall be analyzed, and their validity shall be established.

A quite different projective technique utilizing drawings is the *Draw-a-Person Test*.⁴⁰ Used with individuals two years of age and older, it simply requires the subject to "draw a person." When the first drawing has been completed, he is instructed to draw a person of the opposite sex. Completion of the drawings is followed by an inquiry, the subject being asked to tell a story about each "person" he has drawn as though he or she were a character in a play or a novel. The examiner may also ask a series of prescribed questions about the "characters" in order to obtain added information about environmental factors and relationships.

Each drawing is analyzed regarding certain specified characteristics, and intercomparisons of the two drawings are made in order to discern the subject's attitudes toward himself, and toward his own as well as the opposite sex. In general, analysis and interpretations of drawings are based upon the hypothesis that the drawings represent one's conception of his body in the environment. The drawing of the figure representing one's sex is regarded as a "body image"—a symbol of the concept of the self, a reflection of self-regard. In drawing the human figure of either sex, it is hypothesized, the subject becomes ego-

³⁹ E.g., P. J. Napoli, "Interpretive Aspects of Finger Painting," *Journal of Psychology*, Vol. 23, 1947, pp. 93-132.

⁴⁰ K. Machover, *Personality Projection in the Drawing of the Human Figure: A Method of Personality Investigation*, Springfield, Ill.: C. C. Thomas, 1949. Also by Machover: "Human Figure Drawings of Children," *Journal of Projective Techniques*, Vol. 17, 1953, pp. 85-91.

involved; conflicting needs and tensions are expressed through details and organization of the drawn figures.

The nature of the analysis can be better appreciated by noting the structural features that are taken into account: size, background, exactness, degree of completion, detailing, symmetry, mid-line emphasis, perspective, reinforcements, proportions, placement on the page, theme, shading, erasures. Content of the drawing is also analyzed: individual body parts, clothing, accessories, facial expression, posture. Aesthetic qualities are not considered.

Several body parts are listed below to illustrate some of the interpretive significance attached to them:

head: intellectual aspirations, rational control, fantasy elaboration of the personality;

eyes: uncertainty, paranoid wariness, sexual appeal;

nose: masculinity and assertiveness;

arms and hands: the primary extensive organs; indicators of degree of power, degree of reaching out, extent of environmental manipulation;

bilateral symmetry: degree of obsessiveness-compulsiveness;

body points: exaggerated somatic preoccupation.

Karen Machover herself has used and experimented with this device for a number of years and has made some acute and insightful personality interpretations with it.⁴¹ Other clinical psychologists, who have been instructed by her, have also made effective use of the test. However, the validation of this test as an instrument for general use awaits intensive study on as objective a basis as the materials permit.

PLAY⁴²

Since play is free of the constraints of ordinary adult activity and free of the restraints imposed by adults upon children, it is useful as a projective technique in the study of less apparent aspects of personality. For it is unstructured, provides opportunity for fantasy and

⁴¹ See L. K. Frank et al., *Personality Development in Adolescent Girls*, Society for Research in Child Development, Monograph No. 53, 1951, pp. 89 ff.

⁴² See V. M. Axline, *Play Therapy*, Boston: Houghton Mifflin, 1947; G. R. Bach, *Young Children's Play Fantasies*, Psychological Monographs, Vol. 59, No. 2, 1945; E. H. Erikson, *Studies in the Interpretation of Play: Clinical Observations of Play Disruption in Young Children*, Genetic Psychology Monographs, Vol. 22, 1940.

imaginativeness to operate, and gives scope to individuality of expression.

Play, as a method of personality diagnosis and therapy used almost solely with children, was first tried as a substitute for the free-association

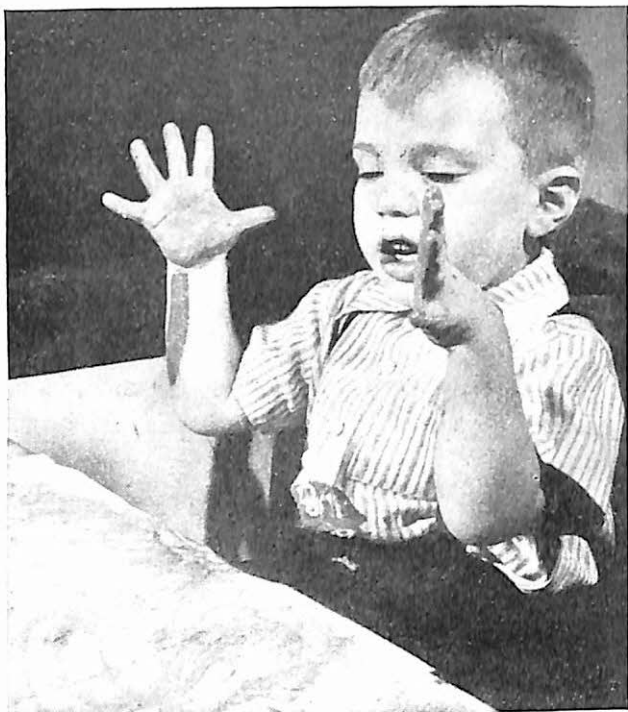


FIG. 20.2. *Finger Painting*. "Sometimes it bleeds! Look!" (Indicates paper painted red.) "Look! Bloody! Like my throat!"

"He smears his hands and arms through the finger paint. And as he pins his thoughts and feelings down on paper he feels, perhaps, more secure. After he has captured them on paper he can handle them a little better." From V. M. Axline, *Play Therapy*, Boston: Houghton Mifflin. (By permission.)

tion technique of psychoanalysis. From such use, investigators have emerged with the theory that play activity is determined by a complex of factors; that it provides an outlet for release of emotional tensions and for overt or symbolic behavior expressing needs, wishes,

desire for experience, attitudes, without fear of censure or punishment.

The commonly used play technique of personality study and therapy consists of this: the child is introduced to a collection of toys which he is permitted to use freely, while the observer notes his activities with respect to the particular items employed, the use made of



FIG. 20.3. Doll Play. She began to play with the family of dolls. "This is Tom," she said to the therapist. "Tom is a funny boy. Want to know what happens to Tom?" From V. M. Axline, *Play Therapy*, Boston: Houghton Mifflin. (By permission.)

them, the organization or patterning of toys, attitudes toward each toy, vocalizations, and general behavior in the play situation. The toys themselves must be such as to provide insights into a child's personality and needs. The items include dolls representing members of the family, furniture (especially kitchen and bathroom appliances), water, sand, vehicles, animals, building blocks, balloons, sticks, or any other pieces that might be relevant in a particular instance.

With play, as with drawing and painting, methods of analysis and interpretation of behavior differ according to the theoretical framework and assumptions of the investigator or therapist. However,

while the play technique of personality study is at present a highly subjective method, numerous case reports attest to its value as a diagnostic and therapeutic procedure.

An effort has been made to standardize and objectify the evaluation of children's play in a therapeutic situation.⁴³ This device, called the World-test, utilizes numerous miniature pieces that represent a large variety of objects and persons found in children's environments. The child is simply instructed to create any setting or situation he might choose. The child's performance and emotional condition are interpreted in terms of the number and variety of pieces used, the structure of the situation (rigid, flexible organization, chaotic), aggressiveness, etc., (depending upon which pieces are used and how they are used). Although this technique is more formalized and has some objective aspects, the basic value of the method depends upon the interpretation of the child's performance and upon demonstrable relationships between this play activity and children's problems of adjustment.

EVALUATION OF PROJECTIVE TESTS

Since the Rorschach and the Murray TAT are the two most widely used projective tests and have been subjected to the most extensive and thorough research, they have been described and explained in some detail. The other projective tests, described and briefly commented on in this chapter, were presented for these reasons: to acquaint the student more fully with psychological thinking in this area, to show how this thinking is now being implemented, and to indicate the variety of instruments.

The following paragraphs summarize the present status and the major problems of current projective tests.

The Rorschach and the TAT have proved to be of considerable value, even though much scientific research (clinical and experimental) remains to be done to fulfill all requirements of a soundly standardized test. It should be recalled, however, that many specialists maintain that the usual specifications of standardization cannot and should not be applied to projective tests.

⁴³ C. Buhler et al., "World-test Standardization Studies," *Journal of Child Psychiatry*, Vol. 2, 1951, pp. 2-81. The original work was published by M. Lowenfeld, in England; "The World Pictures of Children," *British Journal of Medical Psychology*, Vol. 18, 1939, pp. 65-101.

Adequate normative data are lacking for projective devices, even for the Rorschach and the TAT. In some instances no normative data whatever are available. Response norms, both quantitative and qualitative, should be obtained for various groups, according to age, sex, educational levels, and clinical classification. More emphasis upon "normal" populations is needed to provide the proper perspective within which to view performances of clinical groups. The views and interpretations of clinicians have been too strongly affected by their almost exclusive contacts with atypical personalities.

Reliability in terms of agreement among several scorers (sometimes called "scorer reliability") is most significant for these instruments and on the whole has yielded favorable results. Test-retest reliabilities are of limited value and present special difficulties, as pointed out. The method of split-half reliability is inappropriate for reasons indicated.

Validity studies of the matching type have proved most satisfactory. The use of known groups has been only moderately satisfactory because the criteria have been qualitatively and, to some degree, differently defined. Furthermore, the application of these definitions, for purposes of classification, is a subjective matter. The relative unreliability of psychiatric classification has already been noted. Validity studies are often complicated by the fact that interpretation of test data is "global" rather than in terms of limited and measurable elements. Laboratory studies of experimentally induced states or "sets" may be seriously questioned as a method; for they do not deal with basic and established personality traits. A possible exception is the use of hypnosis and similar means.

Although instructions for and procedures in administering a given test do not vary radically among specialists, uniformity is highly desirable. In this connection, the attitude of examiners (reassuring or neutral) is important.

Specialists will have to develop more nearly uniform and objective scoring systems. Since interpretation must be based upon a theory and an understanding of the dynamics of human behavior, markedly different theories lead to varied interpretations and to confusion.

Definitive experimental research remains to be done on the nature of the psychological processes involved in projective tests.

If rapport is established with the examiner, the subject is thereby encouraged to verbalize and understand his behavior, attitudes, values, and the environmental forces.

Appropriately scaled projective techniques are of more interest to subjects than are personality inventories.

Malingering and falsification are more difficult than on personality inventories, though not impossible.

Projective methods—particularly the Rorschach and the TAT—are frequently used by psychologists to supplement information de-

rived from other tests (e.g., intelligence) in order better to understand the complex of factors operative in behavior in a given instance or at a given time.

Projective methods alone are not the answer to all questions regarding human personality and adjustment—a claim made for them by some extreme enthusiasts. But in the hands of a qualified examiner these instruments assist in obtaining information not otherwise available, except possibly through extended psychological interview and observation.

It is apparent that interest in projective methods is widespread among psychologists, from the viewpoints of experimentalists, personality theorists, and clinicians ⁴⁴ dealing with individuals presenting every gradation and variety of problem (not necessarily maladjustment) from childhood through adulthood. It is to be expected, therefore, that solutions of the problems of projective methods will, with time, be more nearly achieved and inadequacies of the instruments reduced.

If the present trend continues, we may expect to have a number of separate projective tests—especially of the thematic picture type—designed for specific and limited age levels, providing relationships and situations typical of and significant for that group. The large crop of new devices, among them being some that are promising and insightful, will be “shaken down”; and the “chaff” will be separated from the “wheat.”

⁴⁴ When we refer to clinicians, we mean not only the psychologist who works with patients in a hospital or in a clinic for persons having serious behavior and adjustment difficulties; we mean also the psychologist who works with any person in an effort to deal with and solve any type of behavioral or adjustment situation through the use of available sound psychological techniques. In other words, a clinician may be one who is concerned with persons within the normal range, as well as one who restricts his activities to the “sick” or “near-sick.”

21.

SITUATIONAL TESTS

AMONG the more recent developments in psychological testing are the situational tests which either test the individual in action or confront him with situations related to his own life, in response to which he gives expression to his feelings for other persons. Although situational tests are not as unstructured as the Rorschach and the TAT, they are in a degree projective methods; for the subject by means of them reveals some of his personality traits through his preference for or against certain contacts with others (as in sociometric tests), and through his spontaneous methods of dealing with life situations, pre-conceived by the examiner, that confront him (as in the psychodrama and in Office of Strategic Services tests).

SOCIOMETRIC METHODS

Description. This method, credited to J. L. Moreno as the innovator,¹ may be defined as a technique for revealing and evaluating the social structure of a group through the measurement of the frequency of acceptance or nonacceptance between the individuals who constitute the group. It is an approach to the problem of studying interpersonal relationships. This technique permits the analysis of each person's position and status within the group, with respect to a particular criterion. (E.g.: Name the pupil in your class with whom you would most like to sit at lunch; name your second choice. Name the

¹ *Who Shall Survive?* New York: Beacon House, 1934.

two persons in your class, in order of preference, whom you would choose as leader on a trip.) The method also reveals the organization of the group, as well as identifying dominant individuals, cliques, cleavages (sex, racial, economic, etc.), and patterns of social attraction and rejection. The reasons for the existing patterns of attraction and repulsion can then be determined if the personality traits of each individual are known and the values of the group as a whole established.

The method is a very simple one. The sociometric test requires that each individual in a given group choose one or more other persons in that group for a specified purpose. In a schoolroom, the pupils may be asked to name their first and second preferences next to whom they wish to sit, or with whom they wish to attend the movies, or with whom they would like to work on a project. Or they may be asked to name one or more individuals in the group who possess certain specified traits; such as the opposites "talkative-silent," "neat-unkempt," etc. Sociometric tests were used in a state training school for girls to determine with whom each individual would prefer to live or work, and with whom each would not want to live or work.² The sociometric method was adopted for use in the armed forces in an effort to identify individuals for specific assignments requiring, for example, leadership and dependability. Thus each individual is viewed in his social relationship to the whole group.

It is apparent that a sociometric test may be devised for innumerable groups and situations. The guiding principles are that each one must be relevant to a life situation of the group, and the items or questions must be such as to require each person in the group to make one or more definite selections revealing certain personal affinities or rejections, or certain values.

As an illustration, we take a class of seventeen pupils—seven girls and ten boys—in a school grade. They are asked to name the two pupils with whom they would prefer to sit at lunch. After the information is obtained, a sociogram is constructed. Of the several kinds of sociograms that have been suggested, the one shown in Figure 21.1 commends itself because it is simply made and easy to interpret. It is known as the "target technique," having been described by North-

² H. H. Jennings, "A Sociometric Study of Emotional and Social Expansiveness," Chapter 30, in R. G. Barker, et al., *Child Behavior and Development*, New York: McGraw-Hill, 1943.

way.³ There are four concentric circles; acceptability scores, based on total number of choices received by each person, are divided into four groups; the lowest quarter is on the outside of the target and the highest are in the center. Each individual is represented on the target according to his acceptability score. The arrows show which individuals have been selected by whom. Solid and broken lines indicate

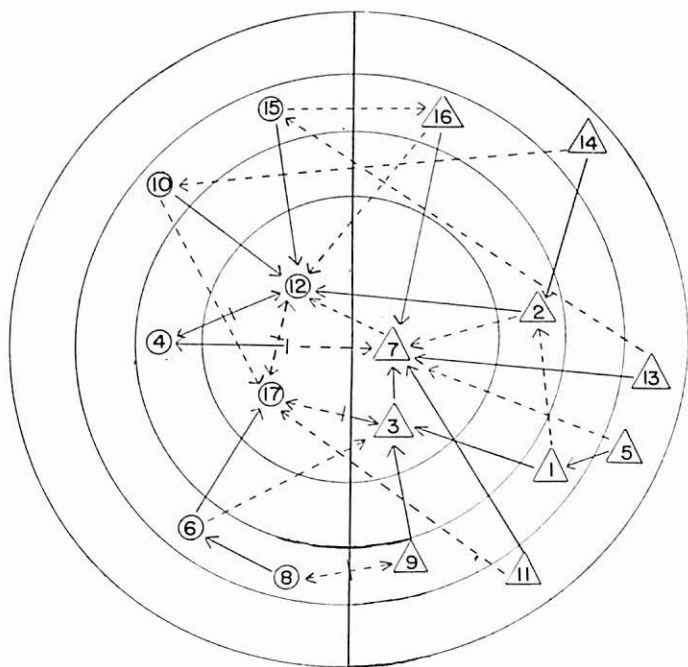


FIG. 21.1. Sociogram of an elementary school class.

first and second choices, respectively. It is possible, also, to divide the target in various ways in order to show cleavages. In this figure, the vertical line readily shows the intersex choices.

Usually, an individual's sociometric score is simply the number of mentions he receives, or the percentage of mentions he receives from others in the group.

Another form of sociometry, which has been called an opinion test, is used to obtain individuals' opinions of one another with re-

³ M. L. Northway, "A Method for Depicting Social Relationships by Sociometric Testing," *Sociometry*, Vol. 3, 1940, pp. 144-150.

gard to a number of traits. The test consists of a series of very brief "word pictures," each one followed by a blank space in which the child is to write the names of others in the group (e.g. classmates) who may be like the "word picture." One may include his own name if he believes the description suits him. Two items are used to describe the extremes of each trait. For example: "Here is someone who finds it hard to sit still in class; he (or she) moves around in his (or her) seat or gets up and walks around." "Here is someone who can work quietly without moving around in his (or her) seat."

In the research reported by Tryon,⁴ twenty pairs of items were used, describing the extremes of twenty traits, which follow:

restless—quiet
 talkative—silent
 attention-getting—non-attention getting
 bossy—submissive
 unkempt—tidy
 fights—avoids fights
 daring—afraid
 leader—follower
 active in games—sedentary
 humor (regarding self)—humorless (regarding self)
 friendly—unfriendly
 popular—unpopular
 good-looking—not good-looking
 enthusiastic—listless
 happy—unhappy
 humor (jokes)—humorless (jokes)
 assured (with adults)—shy (with adults)
 assured (in class)—embarrassed (in class)
 grown-up—childish
 older friends (preference for)—younger friends (preference for)

An individual's score on a given trait is determined by the number of times he is mentioned by his classmates on the pair of opposed items. The item in each pair designating activity is given a positive score, whereas the opposed item—inactivity—is given a negative value. An individual's score on each trait is the algebraic sum of "positive" and "negative" mentions received. The algebraic sum of mentions received by each child on each trait is converted into a pro-

⁴ C. M. Tryon, *Evaluation of Adolescent Personality by Adolescents*, Society for Research in Child Development Monograph, Vol. 4, no. 41, 1939.

portion of the class voting for him in order to equate for the size of the group. (Self-mentions are not included.)

Validity and Reliability. The ordinary questions have arisen in regard to reliability and validity of these sociometric devices. Correlational data indicate that mutual ratings and reratings over short intervals of several weeks, are highly reliable, yielding coefficients of about .90.⁵ Tryon obtained on each trait for each subject two sets of scores by taking the score for one half of the judges (selected at random) and correlating these with the scores of the other half. This split-half method, for the twenty traits, yielded an average reliability coefficient of about .75. Using the test-retest method (ten-day interval), the average correlation between the two sets of scores was about .80. Tryon also reports that only about ten percent of the children and adolescents in her group received disagreeing votes on any pair of items; and most of these individuals were close to the average category in the number of votes received. Regarding individuals who clearly deviated from the average category, however, there was rarely any disagreement. The foregoing evidence indicates that, among the subjects tested, there is a high degree of consistency or concurrence of opinion regarding one another.

The usual criteria and standards of validity do not apply to sociometric tests; for they do not set out to determine what are some of the actual personality and behavior traits of the individuals being rated. They are, rather, measures of the *environment of opinion* in which each individual is functioning. When children and adolescents express preferences for or rejection of classmates, or when they mention classmates in connection with specific traits, they are not necessarily giving their own independent judgments. As a member of the group, each individual acquires, in some degree, the prevailing group attitudes toward his fellows. And as the object of these attitudes, each individual interacts in some manner with these opinions and the persons holding them. Since the purpose of the sociometric technique is not to measure the personality of each individual but to measure the environment of opinion in which he lives, it is gratuitous to ask for or expect evidence of validity in the usual psychometric terms. The discovery of the environment of opinion in which an indi-

⁵ E.g., W. I. Newstetter, et al., *Group Adjustment: A Study in Experimental Sociology*, Cleveland: School of Applied Social Sciences, Western Reserve University, 1938.

vidual lives is none the less important; for such information helps explain that person's behavior as well as the organization of the group and its values.

Uses. In analyzing a group structure by sociometric means, Moreno and others have used the number of "isolates," "mutuals," "unreciprocated choices," and "stars" as indices of group coherence. Frequencies of intersex, interracial, international, inter-occupational-level choices may be used as evidence of cleavages.

Sociometric technique has been applied to the study of a variety of social situations: in classrooms, factories, camps, fraternities, and residential communities. These investigations have revealed the structures of social groups and have provided the bases upon which therapeutic procedures were developed or reconstruction undertaken.

TESTS OF "SOCIAL INTELLIGENCE" AND LEADERSHIP

In the armed forces, during World War II, a number of situational tests were devised to aid in the selection of personnel for assignment to tasks involving leadership and interpersonal relationships.⁶ Two illustrations will be given.

A "Social Manipulation Inventory" describes a variety of problem situations for each of which the examinee is required to indicate the most desirable solution. For example:

You are a supervisor of an office force of 10 people. One member is habitually late. You would:

- A. Make an example of him by discharging him.
- B. Bawl him out in front of the whole group.
- C. Call him in and try to find out the reason for the tardiness.
- D. Call a meeting of the office force to explain that everyone owes it to the company to be on time.
- E. Call him in privately for a lecture on the importance of being on time.

Statistical results for this inventory showed a substantial positive correlation (.30 — .50) with evidence of academic intelligence, thus suggesting that performance scores on the inventory reflect to some extent the subject's knowledge of the expected behavior, but not necessarily his actual behavior in the specified situation.

⁶ J. P. Guilford et al., *Printed Classification Tests*, Report No. 5, U. S. Government Printing Office, 1947, pp. 713 ff.

The number and variety of problem situations in any such situational inventory will depend upon the nature of the situation and the general purpose for which it is being devised.

A second illustration is the "Pilot Behavior Blank," devised to provide information on types of leadership. It was based upon the conception that leadership may be classified as one of three kinds, as divided by Kurt Lewin: laissez-faire (giving help only when requested to do so), authoritarian (dictatorial, domineering), and democratic (participating in a group on a peer basis).⁷

This behavior blank was intended to reveal the kind of social interaction preferred by each subject, on the basis of which combat crews of similar preferences could be assembled. An effort was made to include in this preference blank an equal number of each of the three types of solutions to situations requiring leadership. The items were of the two-alternative kind, the examinee being required to indicate which procedure he prefers. For example:

- A. The pilot who gives many instructions.
- B. The pilot who, while he works more energetically than the rest of the crew, doesn't expect them to work as hard as he.
- A. The pilot who lets the crew members make their own arrangements for quarters, mess, and entertainment.
- B. The pilot who talks about the crew's good points to others.

Although neither one of the foregoing situation-inventories was actually validated, they did yield results—particularly the second one—such as to suggest that they have possibilities for development into instruments of value in assessing and predicting behavior.⁸

PSYCHODRAMA

Aside from a degree of mysticism and extravagant claims made for it, the psychodrama is a useful projective method in the study of personality and in psychotherapy. The technique is one whereby an individual is required to play *spontaneously* an assigned role in a specified situation. The "drama" may involve two or more

⁷ K. Lewin et al., "Patterns of Aggressive Behavior in Experimentally Created Social Climates," *Journal of Social Psychology*, Vol. 10, 1939, pp. 271-299.

⁸ See J. Mathews, *Research on the Development of Valid Situational Tests of Leadership: 1. Survey of the Literature*, Pittsburgh: American Institute for Research, 1951.

persons; it deals with a situation significant in the life of one or more of the participants; each individual may play a role representing either himself or some other personality with whom he is involved.⁹

The central principle of the psychodrama is *spontaneity*, which has been defined by Moreno as the ability of the subject to meet each new situation with adequacy, as "the most important vitalizer of living structure."¹⁰ "The creative idea itself . . . is spontaneous and the quality which pertains to the conception and materialization of such an idea is called spontaneity. Spontaneity must always occur as the first step towards the formation of a cultural conserve."¹¹ In contrast to the act of spontaneity stands the "cultural conserve," which is the creative idea that has become preserved and static, and hence repeated and stereotyped. What the psychodrama aims at, therefore, is to develop in the subject the capacity to play his life roles in a spontaneous and always creative manner which enables him adequately to meet the demands of new and evolving situations, rather than by employing stereotyped patterns of response.

The psychodrama involves a *director*, who is the therapist or the one studying the personalities in the situation. On the basis of knowledge of his subjects and their problems, he creates the situation, selects the actors, assigns roles, observes and interprets the action, and acts as the link between actors and audience. Emphatic and active participation by the audience is an essential of this technique; for its members are individuals who are or will be in situations similar to those being portrayed in the act. The individual who is the subject in the drama (or the patient, in therapy) is called the *primary ego*. He is the one being assisted in the solution of a problem of adjustment, or in learning to live a certain role in life. The *auxiliary ego* is another actor in the drama; he is the agent who provides the assistance needed by the primary ego. The auxiliary ego does so either by (1) acting as the primary ego, identifying with him and representing him toward others or (2) by acting in the role of and representing another person with whom the primary ego is involved.

The foregoing are the basic procedures. A number of modifications and variations of the technique have been evolved. For example,

⁹ See J. L. Moreno, *Psychodrama*, Vol. 1, New York: Beacon House, 1946.

¹⁰ *Ibid.*, p. 100.

¹¹ *Ibid.*, p. 123.

where the two persons involved in a conflict (e.g., parent and child, husband and wife, employer and employee) are placed in a psychodramatic situation, each may be instructed to act out his own role, or each may be required to act the role of the *other* person, as he perceives that other one in the specified situation.

Briefly stated, the psychological rationale of the psychodrama is the following. In therapy, the subject by acting, by participating, in the reproduction of a life situation significant to him experiences an emotional catharsis.¹² In the process, while he gains insight into his own behavior, he is also learning how to meet a situation adequately (spontaneously and creatively) through observations of himself and through interpretations and evaluations given by the therapist (or director) and members of the audience. The psychodrama, thus, is intended to be a *learning procedure* which will teach the subject how to meet life situations in a manner adequate to each one as it arises. Moreno and adherents to his theory and practices report that group psychodrama is especially valuable in the treatment of minor maladjustments, incipient neuroses, and less severe interpersonal conflicts. In the more severe cases of maladjustment and behavior disorder, the group procedure is regarded as a first step toward orientation and personalized (or individual) psychodrama treatment.

In personality evaluation by means of psychodrama, the director and others are able to observe and analyze each subject's characteristic ways of dealing with a situation commonly encountered. A very wide range of themes may be used for testing behavior, depending upon the nature of the persons involved: economic problems, family relationships, social status, school status, play activities, levels of aspiration and self-realization, etc. The number and kinds of themes are as varied as life itself.

The psychodramatic technique, it has been urged, can also be used in vocational selection and training. The recommended method is that each subject be placed in one or more situations created to simulate a particular occupation and to observe his behavior and attitudes as the worker. The individual may then be asked to reverse roles. For

¹² Adherents of the psychodramatic technique maintain that the catharsis derived from it is much more genuine than catharsis derived from the method of psychoanalysis in which the subject merely verbalizes in a remote situation (the therapist's office), dissociated from other persons involved in the conflict.

instance, in selecting store clerks, each subject might be required to take the role first of salesperson and then of customer. Similar arrangements might be made for teacher-pupil, nurse-patient, foreman-worker, waitress-customer relationships, and many others in which interpersonal relationships are an essential part of the occupation.

Moreno has analyzed spontaneity into four characteristic forms of expression, which follow.¹³

(1) Spontaneity which *activates* cultural conserves and social stereotypes. In this form of expression there is nothing creative or original. It is repetition of stereotyped behavior, the only spontaneous element in it being a newness or freshness of feeling and vivacity, that is, its dramatic quality.

(2) Spontaneity which *creates* new forms of behavior and art, and new patterns of environment. Individuals having this quality are rich in creative ideas; they analyze and break up existing stereotypes; they endeavor to produce novel ideas, experiences, and objects. A highly spontaneous person will use his mental resources most effectively and will surpass intellectually more able individuals who do not possess the necessary spontaneity. [This view of Moreno's is an oversimplification and very inadequate explanation of individual differences in achievement.]

(3) Spontaneity which *originates* but is not significant and unique enough to be creative. This form is an expansion on and variation of the cultural stereotype used as a model.

(4) Spontaneity which is *adequate* or appropriate as a response to a new situation. A person might be dramatic, creative, or original; but his responses might, at the same time, be inappropriate to the situation at hand. Adequacy of spontaneity has been called "plastic adaptation skill" essential to "a rapidly growing organism in a rapidly changing environment."

To test an individual's degree and kinds of spontaneity of behavior—which is said to be a fundamental property of his personality—he is placed in life situations where his actions are observed and analyzed. In any psychodramatic situation, analysis of responses may be made within four categories: imaginal content, methods of perception, field-involvement and organization (i.e., kinds of utilization and organization of what the subject has to work with), and social interaction. Analyses and interpretations are very largely qualitative and subjective.

¹³ Moreno, *op. cit.*, p. 89 ff.

An effort, however, has been made to evaluate the roles an individual assumes, on the basis of eight characteristics: ¹⁴

- (1) origin: collective or individual roles
- (2) degree of freedom of spontaneity: role-taking or role-creating
- (3) content: psychosomatic, dramatic, or social roles
- (4) quantity: adequacy or deficiency of roles, superiority or inferiority of roles
- (5) time: slow, average, fast, or "overheated" in developing a role
- (6) consistency: weak, balanced, or strong
- (7) rank: dominant or recessive
- (8) form: flexible or rigid

The kinds of roles an individual assumes, under these headings, are interpreted as being expressions of his personality. This method is by no means a quantitative measure; but it may, at least, provide the beginnings of an orderly analysis of behavior, which is essential if the psychodramatic method is to be employed on a wide scale.

As already indicated, this technique may be used for personality diagnosis and for purposes of psychotherapy. It may be employed as an individual device, in which only therapist and subject are involved, or as a group technique in which the group may vary in size from three persons (therapist and two subjects) to as many as the director-therapist believes he can work with (performers and audience).

A variant of the psychodrama is called the *sociodrama*.¹⁵ The latter differs from the former in respect to purpose and emphasis. Whereas the psychodrama deals with interpersonal relations and adjustment problems within the individual, the sociodrama is concerned with group values and group structure and thinking. The sociodrama portrays *social* phenomena and conflicts with which the audience is concerned and to which a solution is being sought.

The psychodrama is at the present time a technique still in its early stages, yet to be developed into a device of demonstrated diagnostic value for wide use. In spite of the fact that it is the kind of technique in the use of which a psychologist exercises a high degree

¹⁴ J. L. and F. B. Moreno, "Spontaneity Theory of Child Development," *Sociometry*, Vol. 7, 1944, pp. 89-128. For a modification and elaboration of Psychodrama methods, see J. Del Toro and P. Corneytz, "Psychodrama as Expressive and Projective Technique," *Sociometry*, Vol. 8, 1944, pp. 356-375.

¹⁵ J. L. Moreno, "The Concept of Sociodrama," *Sociometry*, Vol. 6, 1943, pp. 434-449; also *Psychodrama*, p. 315 ff.

of personal judgment and evaluation, before it can be widely applied and taught it must achieve a higher level of standardization in respect to techniques of observation, rating of activity, and interpretation of responses. Furthermore, the hypotheses regarding the value of the psychodrama as a technique for developing spontaneous, adequate, and adjusted personalities has not been sufficiently demonstrated. We may expect, however, that the serious students who have interested themselves in this method will provide evidence on this score. Finally, one serious obstacle to the experimental development and use of the psychodramatic group technique is its heavy requirements of time, personnel, facilities, and equipment. These considerations will prove to be a handicap; so we may expect further research to come only from a few sources.

OFFICE OF STRATEGIC SERVICES: ASSESSMENT TESTS ¹⁶

Description and Procedure. During the late war, a group of psychologists and psychiatrists were given the assignment of assessing the traits of men and women recruited for the O.S.S., as it has come to be known. The task was to devise test procedures that would reveal the recruits' personalities and give reliable predictions of their future usefulness in this branch of military service. At the main station, the testing period lasted three days; at another station it lasted one day.

The number and types of jobs in this organization were large and varied, including script writer, base-station operator, demolitions instructor, field representative, section leader, resistance-group leader, saboteur, undercover agent, liaison pilot, pigeonier, and others. Very little indeed was known regarding the specific qualifications essential for successful performance in the jobs to be filled. Nor was there time and opportunity to proceed with job analyses, as would be done in the case of an ordinary civilian situation when aptitude and personality tests for a specific occupation are to be devised. The assessment staff decided, therefore, to use the "organismic" approach: that is, to evaluate each personality as a whole. This meant that some members of the staff would provide an over-all evaluation and description of each individual, based upon interview; each candidate would be tested, observed and evaluated in respect to specific traits

¹⁶ O.S.S. Staff, *Assessment of Men*, New York: Rinehart, 1948. The idea for the organization of the assessment work and some of the procedures used was gained from the British War Office Selection Board.

of personality, intellect, and physique; and finally, all information for each individual would be assembled, organized, and interrelated to provide a complete description and evaluation of each candidate. On the basis of their unified conception of each individual's personality traits, the staff estimated the probable level of future performance. For each recruit, then, a proposed assignment was determined upon, using as criteria the statements of the qualifications required for each job as formulated by each branch of the O.S.S.

Among the variety of devices used in the assessment of the candidates were situational tests which were very similar in conception to those used in the psychodrama and sociodrama. A few of them are listed below:

Upon arrival at the testing station, each candidate was judged according to the ease with which he used the fictitious name under which he went (candidates did not know each other's real names, ranks, or civilian status); physical agility in getting off the truck.

The first day, during the welcoming talk, each candidate's attitudes, postures, questions, and comments were noted.

During the first meal, each recruit's conversation was noted (topics revealing identity were prohibited), as was ease of establishing contact with others.

Various other observations were made during the first evening, in the free periods, when the situations were relatively unstructured.

Terrain test: Candidates were told that at noon of the following day they would be tested for ability to observe the terrain of the station and its buildings, and from their observations, to infer the history of the farm.

Crossing a brook: The task was for a group to carry a delicate instrument over a "raging torrent" with "sheer banks" and to return with some material from the other side. Available were a few boards, a log, a heavy rock, lengths of rope, a pulley, and a barrel with both ends knocked out. All members of the group were on an equal footing, no one having been designated as leader. The actual setting was a shallow, quiet stream, about eight feet wide.

Construction: Ostensibly a test of the candidate's ability to direct two helpers in building with him a frame structure out of simple wooden materials. Actually it was a test of leadership, emotional stability, and frustration tolerance, for the two helpers were junior staff members. The job of one was to act passive, sluggish, and even as an obstacle. The job of the other was to be aggressive, offer poor suggestions, express dissatisfaction and criticism.

Interrogation test: Candidates selected for intelligence work were tested for skill in eliciting information in an interview and in using

the information gained. A military situation was simulated, such as interrogating an American soldier who has escaped from an enemy camp, in order to obtain information about the enemy.

Stress interviews: This situation was devised to test the candidate's capacity to tolerate severe emotional and intellectual strain. Strain was created by rapid, confusing cross-questioning under disagreeable conditions with the aim of detecting flaws in a covering-up story which the candidate has been given twelve minutes to invent.

Post-stress interview: Following the stress interview, each candidate, individually, was placed in a relaxed atmosphere, in the presence of a staff member. Casual conversation was started by the latter in an effort to get the candidate to "let down his guard" and thus reveal information about himself and his experiences at the station that should have been kept to himself.

Assigned leadership: Field problems were devised; in each the group would work as a team, with one member assigned the role of leader. This was used as a test of planning ability and skill in meeting emergencies.

Improvisations: For each candidate appropriate dramatic situations were invented, to which he had to respond realistically in a psychodrama. These improvisations were used to test out and clarify "critical hunches" regarding certain personality traits of each candidate. Leadership quality, moods, attitudes, and modes of conducting interpersonal relationships were evaluated by this method.

After interviewing branch chiefs to obtain their notions regarding the necessary traits for successful performance, and after organizing and combining the listed traits, seven major variables emerged; and to evaluate these in each candidate, the several kinds of tests were devised, of which those described above are samples.

The seven general variables, considered to be basic to the needs of the O.S.S., were:

- (1) motivation for assignment: war morale, interest in proposed job
- (2) energy and initiative: activity level, zest, effort, initiative
- (3) effective intelligence: practical and efficient utilization of intelligence in dealing with things, people, and ideas
- (4) emotional stability: steadiness, endurance, control over disturbing emotions, freedom from neurotic tendencies
- (5) social relations: good will, team work, freedom from disturbing prejudices and annoying traits
- (6) leadership: initiative, ability to evoke cooperation, to organize, administer, and accept responsibility
- (7) security: ability to keep secrets, caution, discretion, ability to bluff and mislead

To these seven traits, three others were subsequently added for use in selected jobs requiring them, namely:

- (8) physical ability: agility, daring, ruggedness, stamina
- (9) observing and reporting: observation and accurate recall of significant facts and their relations, evaluation and succinct reporting of information
- (10) propaganda skills: ability to perceive the psychological vulnerability of the enemy, to devise subversive techniques, and to speak, write, or draw persuasively

As each candidate was observed in the several individual and group situations, he was rated independently by staff members, for all the variables involved in a given test, on what was in fact a sixteen-point scale, although there were only six categories, thus:

- 0. very inferior
- 1. inferior
- 2. low average
- 3. high average
- 4. superior
- 5. very superior

Since the judges found it desirable and possible to add plus or minus ratings to these numbers (omitting scores of 0— and 5+), they were actually using a sixteen-point rating scale. The team of staff members later met to pool their judgments and to arrive at a final combined rating for a given test situation.

The final step in the assessment program was a staff meeting in which the complete reports on each candidate were presented, evaluated, criticized, in which differences of interpretation and viewpoint were resolved and final disposition of each recruit decided on. "The purpose of the meeting was to arrive at an optimal characterization and evaluation of the candidate, and unless the report as presented was felt by the group to be completely adequate, it had to be made so through the cooperative effort of the group. While in the majority of cases the report was accepted with minor changes only, some problematic cases were reviewed and reworked in great detail. This involved another pooling of observations and interpretations."¹⁷ In effect, then, the entire procedure employed situational (and other) test techniques, rating scales, and the case conference (clinical) method.

¹⁷ *Ibid.*, p. 212.

Validity and Reliability. Since the task of the staff was to devise tests that would reveal personality traits for the purpose of predicting success in future assignments, it was necessary to appraise the forecasting value of the procedures being used. Even in ordinary civilian situations, where subjects are under frequent or constant observation and where their effectiveness in performance can be judged in terms of relatively concrete outcomes, the making of ratings presents some serious difficulties, as we have already noted in an earlier chapter. It was to be expected, therefore, that evaluations of the performance of persons accepted after O.S.S. assessment would be even more difficult and less reliable; for these men and women were not always under close observation in the field; it was not always possible to rate their work, because often the results were intangible and deferred; and, for the most part, the primary judges on the job were inexperienced in making psychological evaluations. The following results should be viewed with these considerations in mind.

Four techniques of appraisal were used: (1) overseas staff appraisal, (2) theatre commander appraisal, (3) reassignment area appraisal, and (4) returnee appraisal.¹⁸ In the first instance, assessment staff members obtained, in Europe, appraisals from the immediate chief or commanding officer of the person under consideration, and from several of his associates if possible. The second method was to obtain ratings and a report from the immediate superior of each individual as he returned from an overseas assignment. The third source of validating information was a reallocation center to which individuals were sent for reassignment after they had completed one tour of duty. There they were interviewed by members of the assessment staff in regard to their previous field experiences and rated on anxiety, dejection, homesickness, irritability and quarrels, alcoholism, psychosomatic symptoms, and strength of complaints. The fourth method, returnee appraisal, consisted of obtaining from O.S.S. returnees their ratings of personnel known to them in their area of operations. In this connection, the technique consisted of check lists, questioning by the interviewer, and undirected personality descriptions.

In all instances, by whatever method obtained, the appraisal information was evaluated on a numerical scale and correlated with

¹⁸ *Ibid.*, p. 397 ff.

over-all *assessment* ratings and with specific trait ratings which had been assigned after the initial testing period at the center. Basically, the validity coefficients depend, among other things, upon the reliability of the ratings. In the case of "returnee appraisal" the mean correlation between ratings provided by informants (former associates) was low, being approximately .35. As a measure of reliability—agreement of ratings in this instance—this coefficient is poor and indicates that there were serious differences between the judgments of informants regarding the same individuals being evaluated. When appraisal ratings obtained by each of the four methods were inter-correlated, the coefficients varied from .46 to .59, the mean being approximately .52. Members of the assessment staff, however, agreed much more closely when they themselves rated the field performance of individuals on the basis of information they had obtained from several sources, about each person. In this instance, the mean reliability coefficient (agreement among staff conferees) was approximately .80.

Validity of the original assessment ratings was calculated by correlating them with appraisal ratings. The obtained coefficients were disappointingly low: .08 and .30 (*r* with reassignment area appraisal); .19 and .21 (*r* with returnee appraisal); .23 and .15 (*r* with theatre commander's comments); .37 and .53 (*r* with overseas staff appraisal).¹⁹ Obviously the appraisal ratings made by staff psychologists yielded the most satisfactory results. This is due, first, to their professional experience and greater competence in evaluating behavior, and, second, to the fact that as members of the staff they were applying the same criteria in the rating of field behavior as they and their colleagues had applied in their original assessment ratings.

It is to be expected that correlations between assessment ratings of each of the separate traits and appraisal ratings will be lower than those already cited for over-all trait ratings, since the latter represents a composite of factors. The coefficients for the separate traits (correlated against each of the methods of appraisal) varied from $-.14$ to $+.53$; the median coefficients (for each trait correlated with the four methods of appraisal) ranged from .06 (social relations) to .32 (effective intelligence).

Statistics of greater significance than correlation coefficients, in

¹⁹ Two coefficients were found for each form of appraisal because the subjects assessed at each of two stations were separately treated.

judging the effectiveness of the assessment procedures, are the percentages of unsatisfactory cases in the field, found among the men and women who had been passed as satisfactory (high or medium) by the assessment staff. The percentages of unsatisfactory individuals from the groups rated high or medium (combined), as found by each of the four types of appraisal were: 14.8 and 6.0 (overseas staff appraisal); 13.4 and 15.2 (theatre commander's comments); 11.3 and 4.5 (reassignment-area appraisal); 16.1 and 3.5 (returnee appraisal).²⁰

Although no data are available with regard to individual failures of personnel prior to the introduction of assessment procedures, it is safe to assume that they were higher, probably much higher, than after assessment. It would be very unusual indeed if this were not the case; for such has been the usual experience when psychological methods, though certainly imperfect, have been introduced where none existed before. In any event, the foregoing percentages of failure are low, on the whole, and suggest the value of continued work and research with situational tests.

The authors of the assessment study suggest that the data on the predictive value of the assessment procedures employed were not more impressive because of some or all of the following reasons: (1) defects of the appraisal methods; (2) defects of the assessment methods; (3) assessment staff's lack of familiarity with jobs to be filled and conditions under which personnel would work; (4) the shifting and unpredictable conditions under which overseas personnel worked. No doubt, each of these was in an undetermined degree responsible for the unimpressive validity coefficients.

Since the termination of the war in 1945, little has been done along the lines of the O.S.S. situational tests, as a result probably of the difficulties and complexities inherent in the method, as already indicated. One similar project, at the University of Michigan, was concerned with the selection of graduate students in clinical psychology.²¹ The subjects studied were 137 men who had already been accepted in a number of universities for preparation in clinical psychology, under the Veterans Administration program. Although this project followed the principles of the O.S.S. program, to some extent

²⁰ Percentages are given separately for the two classification centers.

²¹ E. L. Kelly and D. W. Fiske, *The Prediction of Performance in Clinical Psychology*, Ann Arbor: University of Michigan Press, 1951.

adapted to civilian situations, additional forms of psychological evaluation were used, such as projective tests, personality inventories, achievement and aptitude tests, and interest questionnaires. Also, the period of observation was longer than that of the O.S.S., being nine days as compared with three (or at times only one).

The purpose of the Michigan project was to discover, if possible, the traits that contribute to or detract from competence and success in the practice of clinical psychology. Since there were few adequate criteria beyond academic ratings in graduate studies, the findings were inconclusive. This is not surprising, for the observed behaviors and the personality traits rated in the situational tests (for example, cooperation, group participation, expressive movements and ability to empathize) are not necessarily associated with academic and intellectual aspects of professional preparation. Determination of the actual validity and predictive significance of the situational tests used in the Michigan project will have to wait until the graduate students who were the subjects have had significant professional experience and can be rated by competent judges who will use clearly defined criteria.

EVALUATION OF SITUATIONAL TESTS

The several psychological techniques presented in this chapter are not at present of equal value, nor at comparable levels of development, nor applicable with equal facility. Sociometric methods are farthest along in development, can most readily be applied, and yield results that are most easily interpreted. They are valuable in the hands of professional psychologists and other professional groups, in furnishing descriptions of group structures and of individual status within the group. They do not provide information regarding the causes, or factors, that account for the structure or status. These can be determined only through further and close study of the individuals and the community involved.

The two situation inventories reported are illustrations of the questionnaire technique applied to a specific problem, devised for the purpose of revealing attitudes and preferences in that situation. As such, they should be viewed and evaluated in much the same terms as the personality inventories already described in an earlier chapter.

The psychodrama (and the sociodrama) are based upon the long-recognized value of psychological catharsis, but catharsis through ac-

tivity rather than verbalization. The proponents of this technique also claim that it develops spontaneity of behavior, which promotes wholesome development and adjustment. This remains to be demonstrated. The psychodrama, in addition, is so devised as to provide the subjects with opportunities to gain insight into their conflicts and into the attitudes of other persons involved with them in the conflict. In this respect, the psychological rationale appears to be sound. The test, however, of the validity of any technique of personality diagnosis or therapy is a pragmatic one. And here, while a number of case studies have been reported, we must still reserve final judgment regarding efficacy of the method until fuller statistics are available. Deficiency of adequate statistics is not peculiar to the psychodrama; all therapeutic methods are deficient in this regard.

The situational tests used in the O.S.S. and elsewhere are basically sound, from the psychological viewpoint, in that they demand activity of the subject in a situation which simulates the actual setting or task to be performed. The tests were devised to yield evaluations of specified personality traits. The creation of situational tests to bring out particular traits, their numerical rating on a scale, and the statistical study of their interrelations demonstrate that situational tests and psychodrama, as testing techniques, need not be entirely a matter of subjective and intuitive judgment.

It is apparent, of course, that the use of the situational test technique is, at best, difficult; it often requires elaborate facilities and a staff of experienced psychologists (large or small) who are able to diagnose and interpret behavior is essential. The complexity and difficulty of situational testing is evident from the necessary characteristics of a system of personality assessment, as advocated by the authors of the O.S.S. report, and which follows: ²²

(1) Social setting: the whole program to be conducted within a social matrix of staff and candidates, permitting frequent informal contacts and opportunities to observe typical modes of response to other persons.

(2) Multiform procedures: many different techniques to be employed; standardized tests, uncontrolled situations, performance tests, projectives methods, and interview.

(3) Lifelike tasks: in a lifelike environment; complicated tasks requiring organization of thought at a high integrative level, and

²² *Assessment of Men*, p. 464.

some of them to be performed under stress and in collaboration with others.

(4) Formulations of personality: collection of sufficient data to permit conceptualization of the form of some of the chief components of the personality of each individual; the formulation to be used as the basis in making recommendations and predictions.

(5) Staff conference: interpretations of the behavior of each individual at a final meeting of staff members; ratings and recommendations to be reached by consensus.

(6) Tabulations of assessments: formulations of personality, ratings of traits, and predictions of effectiveness to be recorded for the purpose of statistical treatment and precise comparisons with later appraisals.

(7) Valid appraisal procedures: special attention to be devoted to the perfection of appraisal techniques, to determine the validity of each test in the assessment program and of ratings of each variable.

It is not probable that this ideal program will be achieved even after a long time, or even in a very few centers of personality study. In the meantime, close approximations can be and have been achieved in some psychological clinics, where all the tests can be administered and conditions met, excepting the situational tests. As a substitute for these, psychologists and other qualified persons have made detailed observations and rated behavior of subjects, not in "lifelike" situations but in *actual* life situations: for example, children in the classroom or on the playground, teachers in the classroom, adolescents in clubs and in games, employees at work, man and wife during a discussion. Obviously, there will always be some personality diagnoses and predictions that will have to depend upon situational tests (simulated, or lifelike situations) simply because it is impossible to place and observe the individual in the actual situation and task. For this purpose and for personality study of individuals under controlled conditions,²³ the situational test technique holds much promise.

²³ For example, a subject has unresolved needs for aggression which cause him to respond to situations on the basis of personalities involved rather than in terms of the issue at hand. This hypothesis might be specifically examined in a specially devised situational test of the discussion type. Or personality organization might be probed by placing subjects in unconventional situations of varying degrees of stress, relatively devoid of cultural cues and barriers, so that culturally controlled and stereotyped behavior are much less likely to be manifested.

INDEX

- Ability
 - to educe relations, 144
 - range of, 20
 - to verbalize, 164
- Aborn, M., 138, 175
- Abstract reasoning, 330
- Abstract thinking, 61
- Abstraction, 216
- Abstractness, 64
- Accomplishment quotient, 46
- Achievement age, 380
- Achievement quotient, 46, 381
- Achievement tests, evaluation of, 399
- Adams, W. M., 352, 353
- Adaptation, 60
- Adaptation Board, 199
- Adaptiveness to a goal, 65
- Adult intelligence quotient, 125
- Adult level, 154
- Adult mental age, 125
- Adult scale, need for an, 165
- Age criterion, 131
- Age groups, 371
- Age-scale, 147
- Aikin, W. M., 398
- Algebra, 395
- Allport A-S Reaction Study, 468
- Allport, G. W., 453, 455, 487, 495
- Alster, B., 470
- American Council on Education: Psychological Examination for College Freshmen, 288
- Analogies test, 276, 348, 383
- Analogous functions, 401
- Anderson, G. L., 556
- Anderson, H. H., 556
- Anderson, R. G., 530
- Andrews, G., 558
- Anemia, 55
- Anxiety state, 425
- Aptitude, 7, 306
 - in academic subjects, 395
 - in law, 351
 - mechanical, 31
 - in medicine, 344
 - teaching, 355
 - testing program, 375
 - tests of, 2, 28, 306, 336
- Aptitude Classification Tests, 333
- Aptitude tests, general evaluation of, 372
- Area sampling, 491
- Arithmetic, 89, 164, 299
 - tests of, 392
- Arithmetical reasoning, 159
- Army Alpha Test, 165
- Army Beta Examination, Revised, 165, 248
- Army General Classification Test, 294
- Arnheim, R., 514
- Aronson, M. L., 555
- Art tests, evaluation of, 343
- Arthur, G., 217
- Arthur Point Scale, 192, 205
 - revised, 207
- Artificial language, 292
- Ash, P., 501
- Aspects of Personality, 470

- Assembly Test of General Mechanical Ability, 317
 Association, 68
 Associative processes, 52
 Attention, 54, 68
 Attention span, 216
 Attitudes and values, tests of, 484
 Auditory acuity, 310
 Axline, V. M., 571

 Babcock Test, 432
 Bach, G. R., 571
 Baker, H. J., 327
 Baldwin, A. L., 149, 151
 Balinsky, B., 183
 Balken, E. R., 542
 Baller, W. R., 357
 Barker, R. G., 56, 578
 Bartlett, F. C., 567
 Bayley, N., 237, 239
 Beck, S. J., 505, 513
 Bell Adjustment Inventory, 468
 Bell, J. E., 504
 Bellak, L. B., 536
 Bender Visual-Motor Gestalt Test, 447
 Benjamin, J. D., 525
 Bennett, G. K., 313, 322, 329
 Bernreuter Personality Inventory, 469
 Best answer test, 276
 Biersdorf, K. R., 554
 Bijou, S. W., 412
 Billingslea, F. Y., 447
 Binet, A., 9, 69, 97, 113
 early work of, 99
 major contributions of, 111
 Binet Scale, 96
 1905 version, 101
 1908 version, 105
 1911 version, 109
 early American revisions of, 114
 Biographical data questionnaires, 482
 Birren, J. E., 183
 Bizarre responses, 54
 Blacky Pictures, 555
 Block building test, 216
 Block counting test, 247
 Block design test, 162, 164
 Blocking, 55
 Bloss, P., 567
 Blum, G., 555
 Boas, F., 97
 Boyd, F., 440
 Brodman, K., 475
 Brosin, H. W., 520
 Brower, D., 528
 Brown, A. W., 246
 Buhler, C., 531, 574
 Burchard, K. A., 414
 Burnham, P. S., 291
 Buros, O. K., 183, 337, 386, 562, 564
 Burt, C. L., 86, 144
 Byers, J., 431, 566

 C-score, 228
 California Achievement Tests, 385
 California Test of Personality, 469
 California Tests of Mental Maturity, 270
 Callis, R., 357
 Canter, A. H., 175
 Carl Hollow Square Scale, 214
 Carlson, R., 509
 Carp, A. L., 522
 Case of Mickey Murphy, 357
 Casuist board, 198
 Catharsis, 585
 Cattell Culture-Free Test, 256
 Cattell Developmental and Intelligence Scale, 230
 Cattell, J. McK., 97
 Cattell, P., 230, 234, 239
 Cattell, R. B., 257, 259, 495
 CAVD test, 73, 171
 Chance factors, 11
 Character disorders, 54
 Chave, E. J., 485
 Check lists, 383
 Chicago Nonverbal Examination, 246
 Children's Apperception Test, 553
 Clark, R., 538
 Clark, W. W., 270, 385, 469
 Clarke, H. J., 561
 Classification test, 276, 347
 Clerical accuracy, 330
 Clerical aptitude, tests of, 326
 Clerical speed, 330
 Clerical tests, evaluation of, 328
 Clinical aspects of tests, 51
 Clinical diagnosis, 417
 Cofer, C. N., 410
 Cohen, J., 183
 College Entrance Examination Board:
 Scholastic Aptitude Test, 290

- Color vision, 308
- Combs, A. W., 538, 543
- Comparisons, quantitative, 348
- Completion test, 299, 347, 383
- Complex educational objectives, tests of, 397
- Complex mental processes, 112
- Complexity of tasks, 64
- Comprehension tests, 88, 163
- Compulsiveness, 54
- Comrey, A. L., 483
- Concept formation, 71
 - tests of, 435
- Conrad, H. S., 499
- Content analysis, 414
- Content validation, 399
- Cook, W. W., 357
- Corey, S. M., 488
- Cornell-Coxe Performance Ability Scale, 203
- Cornell Index, 37, 475
- Corneytz, P., 587
- Correlation
 - biserial, 33
 - effect of range on, 202
 - multiple, 34
 - rank-order, 23
 - simple, 31
 - tetrachoric, 33
- Coxe-Orleans Prognosis Test of Teaching Ability, 355
- Crawford, A. B., 291, 353
- Crawford, D. M., 312
- Crawford, J. E., 312
- Creative abilities, 153
- Cruikshank, R. M., 313
- Cube tapping sequences, 216
- Cube test, 199
- Cultural influences, 268
- Cunningham, B. V., 245
- Custom-built tests, 401
- Cut-off score, 36, 477
- Cutts, R. A., 414
- Dangling ring behavior, 222
- Darley, J. G., 471
- David, H. P., 564
- Davidson, H. H., 507, 515
- Davis, F. B., 40
- Davis-Eells Test of General Intelligence, 260
- Davison, A. H., 542
- Decile rank, 44
- Decline in test score, 181
- DeHaan, H., 175
- Delinquents, 412
- Del Toro, J., 587
- Deri, S. K., 563
- Derived indexes, 380
- Derner, G. F., 138, 175
- Deterioration index, 426
- Detroit Clerical Aptitudes Examination, 327
- Deviation IQ, 279
- Diagonal test, 199
- Differential Aptitude Tests, 17, 329
- Differential diagnosis, 68
- Difficulty, level of, 64
- Digit-span test, 51, 89, 164
- Digit-symbol test, 163, 164
- Directions test, 299
- Disabilities, 53
- Diversity of materials, 112
- Doll, E. A., 461
- Doll play, 573
- Dot drawing test, 245
- Douglass, H. R., 353
- Drake Musical Memory Test, 338
- Draw-A-Person Test, 570
- Drawing, 569
- Drive, 63
- Durost, W. N., 245
- Durrell Analysis of Reading Difficulty, 388
- Ebaugh, F. G., 525
- Economy of effort, 64
- Educational achievement, tests of, 8, 28, 377
- Educational age, 43, 380
- Educational quotient, 46, 381
- Edwards, A. L., 488
- Efficiency index, 433
- Elkisch, P., 569
- Ellis, A., 496, 498, 499, 555
- Emotional forces, 66
- Emotional states, 68
- Energy, concentration of, 66
- Engineering and Physical Science Aptitude Test, 358
- Environment, 112
- Erikson, E. H., 571
- Eron, L. D., 538
- Expectancy tables, 34

- Experiments, reports of, 59
 Extrinsic factors, 405
 Eysenck, H. J., 530
- Factor analysis, 6, 72, 82
 Factors, 72
 affecting test performance, 404
 illustrations of, 88
 implications of, 94
 residual, 20
- Farnsworth, Paul R., 337
 Feature profile test, 199
 Feingold, S. N., 259
 Fels Parent Behavior Scales, 462
 Ferguson Form Boards, 211
 Ferguson, G. A., 20
 Fernald, G. M., 197
 Figure analogies test, 289
 Fine arts and professions, 336
 Finger painting, 572
 Fiske, D. W., 594
 Five-figure board, 198
 Flanagan, J. C., 86, 333, 469, 498
 Fleischmann, M., 564
 Fleming, E. E., 561
 Fleming, V. V., 131, 154
 Foreign languages, 396
 Forlano, G., 470
 Form boards, 218
 Fosberg, I. A., 520
 Foster, J. C., 227
 Foulds, G. A., 255
 Four-Picture Test, 556
 Frandsen, A. N., 417
 Frank, L. K., 56, 571
 Freedom in response, 53
 Freeman, E., 308
 French, J. W., 495
 Frenkel-Brunswick, E., 542
 Freud, Sigmund, 553
 Fromm, E. O., 520
 Fry, D. F., 323
 Functional analysis, 406
 Functional unities, 30
 Functions
 sampling of, 4
 tested, 58, 266
 tested in the Bellevue scale, 163
 tested in the Stanford-Binet, 141
- G.G.W.S. Object-Sorting Test, 439
 Galton, F., 96, 547
 Gardner, E. F., 385
 Gardner, R. W., 527
 Garfield, S. L., 538
 Garrett, H. E., 86, 147
 Gates Reading Readiness Test, 390
 Gelb-Goldstein Color Sorting Test, 438
 General ability, 80
 tests of, 27
 General comprehension, 159
 General factor, 74, 267
 in the Binet scales, 142
 Gesell, A., 221, 224
 Gesell Developmental Schedules, 220
 Gibby, R. G., 522
 Gilbert, J. A., 97
 Gilliland, A. R., 239
 Goddard, H. H., 114
 Goldfarb, W., 526
 Goldman, R., 424
 Goldstein, K., 435
 Goldstein-Scheerer Cube Test, 436
 Goldstein-Scheerer Stick Test, 441
 Goldstein Tests, evaluation of, 442
 Goodenough Drawing Test, 259
 Goodenough, F. L., 227, 228
 Gorham, T. J., 353
 Gough, H. G., 475
 Graves Design Judgment Test, 342
 Greene, H. A., 395
 Griffin, C. H., 358
 Group factors, 85
 theory of, 77, 86
 Group scales
 for college freshmen, 288
 evaluation of, 293, 301
 uses of, 304
 Group testing, 67
 Group trends, 21
 Grove, W. R., 212
 Guilford, J. P., 6, 30, 86, 93, 458, 481, 495, 519, 582
 Guilford-Zimmerman Temperament Survey, 478
 Guthrie, G. M., 196
- Haggerty-Olson-Wickman Rating Schedules, 460
 Halpern, F., 148
 Hanfmann, E., 436, 566
 Hanfmann-Kasanin Test, 442
 Harris, A. J., 313

- Harrison, R., 538, 541, 543
 Harrison-Stroud Reading Readiness Test, 390
 Harrower, M. R., 518
 Harrower-Erickson, M. R., 529
 Hartman, A. A., 542
 Hartwell, S. W., 558
 Hathaway, S. R., 472, 474, 475
 Havighurst, R. J., 268
 Healy Picture-Completion Test, 199
 Healy Puzzle A, 199
 Healy, W., 197
 Hearing, tests of, 307
 Heimann, R. A., 532
 Henmon-Nelson Tests of Mental Ability, 171, 300
 Henri, V., 101
 Henry, W. E., 541, 543
 Herring, J. P., 115
 Hertz, M. R., 506, 509, 519, 523, 525, 528
 Higginbotham, S. A., 525
 High school tests, evaluation of, 396
 Higher mental functions, 99
 Hildreth, G., 565
 Hogan, H. P., 428
 Holzberg, J. D., 520, 562
 Holzinger, K. J., 86
 Homographic free association test, 550
 Hopi Indians, 268
 Hotelling, H., 86
 Hoyt, C., 20
 Hunt, H. F., 445, 555
 Hunt, J. McV., 410
 Hunt-Minnesota Test for Organic Brain Damage, 444
 Hunt, W. A., 428
 Hutt, M. L., 417, 448, 522, 532, 558, 564
 Impairment, patterns of, 434
 Incentive, 63
 Individual differences, 99
 Information test, 54, 88, 159, 163, 276
 Ingraham-Clark Diagnostic Reading Tests, 388
 Insight Test, 568
 Institute of Educational Research Intelligence Scale CAVD, 298
 Intelligence
 abstract, 70
 Intelligence (*Continued*)
 concrete, 69
 definition of, 5, 60, 111
 kinds of, 69
 social, 69
 Intelligence quotient, 46, 111
 adjusted means of, 135
 classification of, 124, 140
 corrected CA divisor for, 134
 distribution of, 122, 136, 169
 means of, 169
 qualitative significance of, 47
 quantitative significance of, 47
 standard deviations of, 169
 of the Stanford-Binet, 133
 variability of, 137
 of the Wechsler-Bellevue, 179
 Intelligence tests as clinical instruments, 403
 Interest inventories, 361
 evaluation of, 366
 Interest level, 57
 Interests, stability of, 370
 Internal consistency of tests, 10
 Interval discrimination test, 338
 Intrinsic factors in score changes, 404
 Iowa Legal Aptitude Test, 352
 Iowa Tests of Educational Achievement, 385
 Item administration, order of, 416
 Item analysis, 38
 Item difficulty, 39
 Items, types of, 145, 383
 Jackson, R. W. B., 20
 Jastrow, J., 97
 Jeffery, M., 444
 Jennings, H. H., 578
 John, E., 144
 Jones, L. V., 144
 Juckem, H., 444
 Jung, C. G., 548
 Kaake, N. A., 554
 Kandel, I. L., 351
 Karlin, L., 562
 Kasanin, J., 436
 Kelley, D. M., 505
 Kelley, T. L., 41, 85, 385
 Kellogg, C. E., 248
 Kelly, E. L., 594
 Kennedy, S., 514

- Kenney, K. C., 488
 Kent, G. H., 154, 548
 Kent Series of Emergency Scales, 428
 Kent-Shakow Form Board Series, 212
 Kite, E. S., 100, 105
 Klein, G. S., 514
 Klopfer, B., 505, 507, 515
 Knauber Art Ability Test, 340
 Kraepelin, E., 98
 Krugman, J. I., 194, 521, 526
 Krugman, M., 147
 Kuder, G. F., 20
 Kuder Preference Record, 361
 Kuhlmann-Anderson Tests, 285
 Kuhlmann, F., 114
 Kumin, E., 211
 Kutner, B., 538
- Language test, 267
 Language usage, 331
 La Piere, R. T., 488
 Lapp, C. J., 358
 Lasaga y Travieso, J. I., 536
 Law School Admissions Test, 352
 Leadership, tests of, 582
 Learning
 ability, 60
 effects of, 25
 Leeds, C. H., 357
 Lefever, D. W., 531
 Legal aptitude tests, evaluation of,
 353
 Lewin, K., 583
 Lewinski, R. J., 170
 Lewis, D., 337
 Likert, R., 486
 Lindzey, G., 487, 545
 Loevinger, J., 95
 Loftus, J. J., 470
 Logical selection test, 276
 Long, J. A., 40
 Lorge, I., 183, 252
 Lowenfeld, M., 574
- Machover, K., 570
 MacMurray, D. A., 200
 MacPhee, H. M., 170
 Madden, R., 385
 Madison, T., 338
 Magaret, A., 151, 414
 Make-A-Picture Story Test, 557
 Manikin test, 199
- Manual dexterity, 311
 Manual tests, 311
 Marcuse, F. L., 554
 Mare and Foal Form Board, 198
 Martinez-Arango, C., 536
 Maslow, A. H., 479
 Masserman, J. H., 542
 Matching figures, 247
 Matching test, 383
 Mathews, J., 583
 Maurer, K. M., 227, 228, 239
 Mayman, M., 538
 McCall, W. A., 252
 McCandless, B. R., 412
 McCary, J. L., 562
 McClelland, D. C., 544
 McKinley, J. C., 472, 475
 McLeish, J., 340
 McNamara, W. J., 471
 McNemar, Q., 131, 141, 142, 148,
 149, 276, 493
 Mean scatter, 420
 Mechanical aptitude tests, evaluation
 of, 324
 Mechanical reasoning, 330
 Medical aptitude tests, evaluation of,
 350
 Medical College Admissions Test,
 346, 350
 Meehl, P. E., 444, 474, 475
 Meier Art Judgment Test, 340
 Mellone, M. A., 149
 Melton, A. W., 313
 Memory, 52
 Memory factor, 78
 Memory span for digits, 160
 Mental ability, analyses of, 72
 Mental age, 43, 107, 109, 112, 133,
 134
 Mental defectives, 51, 171, 412
 Mental growth, curve of, 50
 Mental impairment, tests of, 432
 Merrill, M. A., 129, 140, 155
 Merrill-Palmer Scale of Mental Tests,
 235
 Metropolitan Readiness Test, 390
 Meyer, B. T., 512
 Michigan Picture Test, 558
 Miles, W. R., 313
 Miller Analogies Test, 296
 Miller, W. S., 296
 Minnesota Clerical Test, 328

- Minnesota Mechanical Assembly Test, 319
Minnesota Multiphasic Personality Inventory, 472
Minnesota Paper Formboard (Revised), 321
Minnesota Personality Scale, 471
Minnesota Preschool Scale, 227
Minnesota Spatial Relations Test, 320
Minnesota Teacher Attitude Inventory, 357
Mittelman, B., 475
Mixed group scales of mental ability, 270
Monroe Diagnostic Reading Examination, 389
Monroe Reading Aptitude Tests, 391
Mooney Problem Check List, 479
Moore, B. V., 358
Moreno, F. B., 577, 584, 587
Morgan, C. D., 533
Morris, C. B., 202
Morton, N. W., 248
Moss, F. A., 344
Motivation, 66
Motor tests, 311
Muensterberg, H., 98
Mullen, F. A., 428
Multi-factor theory, 73
Multiple-choice items, 383
Munroe, R. L., 509, 530, 531
Murphy, Gardner, 453
Murphy, L. B., 569
Murray, H. A., 459, 533, 542
Mursell, J. L., 337
Musical aptitude tests, 336
 evaluation of, 338
Napoli, P. J., 570
National Intelligence Test, 204
National Teachers Examinations, 356
Nelson, V. L., 237
Newstetter, W. I., 581
Noll, V. H., 499
Nonlanguage Multi-mental Test, 252
Nonverbal group scales, 241
 evaluation of, 265
Norms, 43, 56, 57
Northway, M. L., 579
Number test, 267
Number-checking test, 27
Number facility, 280
Number factor, 78
Number series test, 289
Numerical ability, 330
Nutritional disturbances, 55
Object assembly test, 162, 164, 218
Objectivity, 2
Obsessive-compulsive individual, 425
Obsolete items, 152
Oehrn, A., 98
Office of Strategic Services: Assessment Tests, 588
Ohio State University Psychological Test, 289
Olsen, M., 353
Operational levels, 4
Opinion polling, 489
Opposites test, 276, 346
Organic damage, 54
Organized stock-taking by tests, 372
Originality, 80, 153
Originals, emergence of, 66
Osgood, C. E., 498
Otis Group Intelligence Scale, 301
O'Toole, C. E., 324
Painting, 569
Paper form board test, 247
Paragraph comprehension test, 348
Parkyn, G. W., 154
Parten, M. B., 3
Pascal, G. R., 448
Pastovic, J. J., 196
Paterson, D. G., 320
Pattern Perception Test, 253
Patterns of response, 52
Pegboard test, 27, 319
Penrose, L. S., 253
Perceiving differences, 327
Perceiving similarities, 327
Percentile rank, 44
Perceptual speed, 202
Performance materials, 130
Performance scales, 156, 197
 evaluation of, 216
 factorial analysis of, 202
 functions tested by, 215
Perlman, J. A., 512, 528
Personality
 definition of, 452
 dynamics, 68
 inventories, 3, 466, 493

- Personality (*Continued*)
 rating scales, 452
 structure, 515
 tests, 7, 452, 466, 502, 547, 577
 traits, 29
 Personnel work, 28
 Physical measurement, 48
 Picture arrangement test, 161, 164, 247
 Picture completion boards, 218
 Picture completion test, 161, 164, 245
 Picture sequence test, 247
 Picture tests of personality, 551
 Pilot Behavior Blank, 583
 Pintner-Cunningham Primary Test, 245
 Pintner Nonlanguage Series: Intermediate Test, 250
 Pintner-Paterson Scale of Performance Tests, 198
 Pintner, R., 129, 470
 Piotrowski, Z. A., 520, 526
 Piper, A. H., 395
 Play, 571
 Plier dexterity test, 312
 Point-scale, 147
 Population sample, 2, 58
 Porteus Maze Test, 205
 Posner, R., 562
 Practice, effects of, 25
 Pre-Engineering Ability Test, 360
 Primary mental abilities, 77
 tests of, 279
 Probable error, 179
 Proficiency tests, 37, 401
 Profile chart, 284
 Profiles, 369
 Prognostic Test of Mechanical Abilities, 323
 Progressive Education Association Behavior Description, 462
 Progressive Matrices Tests, 255
 Projection, 502
 Projective methods, 502, 547
 Projective questionnaire, 566
 Projective tests, 3
 evaluation of, 575
 Psychodrama, 583
 Psychological analyses, 144
 Psychological Corporation General Clerical Test, 327
 Psychological test, definition of, 1
 Psychometric pattern, 412
 Psychotic persons, 425
 Pursuit test, 318
 Qualitative analysis, 53
 Quantitative reasoning, 292
 Questionnaire, 490
 Rabin, A. I., 418, 520
 Rafferty, J. E., 565
 Random comments during testing, 54
 Random sampling, 490
 Rapaport, D., 411, 418, 506, 542, 551, 563
 Rating scales, 3, 452
 evaluation of, 464
 Raven, J. C., 253, 255
 Raw score, 42
 Reaction time, 311
 test of, 315
 Reading tests, 387
 diagnostic, 388
 evaluation of, 392
 Reasoning, 6, 280
 Reasoning factor, 78
 Relative rank, 42
 Reliability, 10, 58
 absolute, 12
 coefficient of, 12, 20
 relative, 12
 split-half, 13
 subtest, 25
 test-retest, 12
 Remmers, H. H., 487
 Renaud, H., 542
 Responses, qualitative aspects of, 423
 Retesting, 10
 Rhode, A. R., 564, 565
 Richards, T. W., 237
 Richardson, M. W., 20
 Riess, B. F., 553
 Riggs, M. M., 414
 Rioch, M. J., 526
 Roberts, J. A. F., 149
 Roe, A., 530
 Rorschach Test, 63, 504
 evaluation of, 531
 Rosanoff, A. J., 548
 Rosenzweig Picture-Frustration Study, 560
 Rosenzweig, S., 561
 Rote memory, 280

- Rothney, J. W. M., 532
 Rotter Incomplete Sentences Blank, 565
 Rotter, J. B., 538, 564
 Rubinstein, B. B., 525
 Ruch, G. M., 385
 Russell, J. T., 374

 Saetveit, J. C., 337
 Sample, stratified, 3
 Sampling theory, 85
 Sanderson, H., 528
 Sanderson, M. H., 520
 Sandiford, P., 40
 Sanford, R. N., 539
 Sarason, S. B., 259, 543
 Sargent, H., 532, 567
 Saxe, C. H., 542
 Scale, 101, 112
 factorial analyzed type of, 149
 for infants and preschool children, 220, 237
 global type of, 149
 spiral-omnibus type of, 243
 Scaling, methods of, 484
 Scatter analysis, 54, 409, 419
 Schafer, R., 421
 Scheerer, M., 435
 Schizophrenics, 434
 Schlesinger, H. J., 514
 Schneidman, E., 557
 Scholastic achievement, tests of, 2
 School learning, 151
 Schrader, W. B., 353
 Schwartz, M. M., 562
 Schwartz, E. K., 553
 Science and engineering aptitude tests, evaluation of, 360
 Scientific and engineering aptitudes, tests of, 357
 Scorers, consistency of, 26
 Scott, W. D., 463
 Seashore-Bennett Stenographic Proficiency Test, 402
 Seashore, C. E., 337
 Seashore, H. G., 189, 329
 Seashore Measures of Musical Talent, 9, 27, 336
 Security-Insecurity Inventory, 479
 Seguin Form Board, 198
 Selection ratio, 370
 Sensory capacity, 314
 Sensory process, 68
 Sentence-completion tests, 291, 346, 564
 Shavzin, A. R., 522
 Ship test, 199
 Shotwell, A. M., 239
 Siegel, M. G., 525
 Siipola, E. M., 528
 Similarities test, 88, 160, 164
 Simon, T., 105
 Simple recall, 383
 Sims SCI Occupational Rating Scale, 488
 Sims, V. M., 488
 Singer, J. L., 525
 Situational tests, 577
 evaluation of, 595
 Sloan, W., 414
 Smith, E. R., 398
 Smith, M., 374
 Snellen Chart, 308
 Social intelligence, tests of, 582
 Social Manipulation Inventory, 582
 Social value, 65
 Sociometric methods, 577
 Space factor, 78, 202
 Space relations, 330
 Spache, G., 146
 Spatial perception, 52, 280
 Spatial relations tests, 320
 Spearman-Brown formula, 14, 16
 Spearman, C., 9, 75, 83
 Special abilities, 52
 Specific factor, 84
 Speech, manner of, 54
 Speed, 80
 Spohn, H. E., 525
 Spontaneity, 586
 SRA Youth Inventory, 478
 Stainbrook, E., 527
 Standard error of measurement, 12, 17
 Standard score, 45
 Standards, 56
 Stanford Achievement Tests, 384
 Stanford Revision of the Binet Scale (1916), 115
 criticisms of, 126
 reliability of, 118
 scoring method of, 121
 validity of, 115

- Stanford Revision of the Binet-Simon Scale (1937), 129, 171, 192, 209, 406
 evaluation of, 147
 reliability of, 132
 short scale of, 146
 validation of, 130
- Stanford Scientific Aptitude Test, 358
- Stein, M. I., 542
- Steiner, M. E., 518
- Stenquist, J. L., 317
- Stern, William, 111
- Stevenson, I., 428
- Stoddard, G. D., 64, 155
- Story completion test, 567
- Story telling test, 567
- Stratified sampling, 490
- Strength of grip, 311
- Stromberg Dexterity Test, 313
- Strong Vocational Interest Blank, 362
- Study of Values, 487
- Stuit, D. B., 352, 464, 482
- Stutsman, R., 235
- Subject age, 380
- Substitution test, 199
- Suci, G. J., 498
- Sullivan, E. T., 270
- Sullivan, P. L., 475
- Suttell, B. J., 448
- Swift, J. W., 520, 525
- Symonds, P. M., 396
- Symonds Picture-Study Test, 554
- Synonyms test, 276
- Szondi Test, 563
- Taylor, H. C., 374
- Teaching-aptitude tests, evaluation of, 356
- Terman, E. L., 252
- Terman, L. M., 115, 125, 129, 276, 385, 409
- Terman-McNemar Test of Mental Ability, 275
- Test
 administering a, 57
 content, 68
 design, 68
 factors in selecting a, 57
 scores, interpretation of, 42
 scoring, 57
 standardization, 10
- Tests
 of attitudes and values, 488
 for college level, 394
 of critical thinking, 377
 for high school level, 394
- Tetrad difference, 84
- Tetrad equation, 84
- Thematic Apperception Test, 533
- Thetford, W. N., 509
- Thompson, C. W., 151, 414
- Thompson, H., 221
- Thompson Modification of the TAT, 553
- Thomson, G. H., 85
- Thorndike, E. L., 9, 69
- Thorndike Intelligence Examination for High-School Graduates, 299
- Thorndike, R. L., 41
- Thornton, G. R., 519
- Thorpe, L. P., 469
- Thurstone, L. L., 6, 9, 77, 85, 95, 279, 288, 372, 484, 551
- Thurstone, T. G., 6, 77, 279, 288
- Tiegs, E. W., 270, 385, 469
- Time interval, 24
- Time requirements, 57
- Tomkins, S. S., 536, 544
- Toops, H. A., 289
- Traits
 definition of, 467
 measured by tests, 58
 sampling of, 4
- Travers, R. M. W., 498
- Travis-Johnston Projection Test, 551
- Triangle test, 199
- True differences, 11
- Tryon, C. M., 580
- Tryon, R. C., 86
- Tunks, L. K., 352
- Two-factor theory, 74, 83
- Two-figure board, 198
- Two-hand coordination test, 316
- Tyler, R. W., 398
- Validating objectives, 41
- Validity, 26, 58
 Binet's criteria of, 107
 content, 399
 criteria of, 27, 29
 cross, 31
 face, 31
 factorial, 30

- Validity (*Continued*)
functional, 26
immediate, 29
intermediate, 29
methods of calculating, 31
operational, 26, 30
ultimate, 29
- Values, tests of, 487
- van Lennep, D. J., 556
- Van Wagenen-Dvorak Diagnostic Examination of Reading Abilities, 389
- Van Wagenen, M. J., 227
- Variable, gross, 4
- Variability differences in the Stanford-Binet Scale (1937), 148
- Variance, 19
analysis of, 12, 19
- Variation, coefficients of, 169
- Varon, E. J., 113
- Verbal analogies test, 289
- Verbal factor, 78
- Verbal group scales of mental ability, 270
- Verbal materials, 150
- Verbal meaning, 280
- Verbal reasoning, 330
- Vernon, P. E., 86, 154, 487, 519
- Vineland Social Maturity Scale, 461
- Vision, tests of, 307
- Visual acuity, 308
- Visual-motor skill, 71
- Vocabulary, 54, 88, 160, 164, 299
- Vocabulary index, 432
- Vocabulary scatter, 420
- Voelker, P. H., 327
- Walton, R. E., 558
- Watson-Glaser Test of Critical Thinking, 398
- Webb, W. B., 175
- Wechsler Adult Intelligence Scale (1955), 187
- Wechsler-Bellevue Scale, 156, 417
content of, 157
criticism and evaluation of, 180, 182
reliability of, 174
scoring of, 177
- Wechsler-Bellevue Scale (*Continued*)
special features of, 181
standardization of, 165
validity of, 167
- Wechsler, D., 63, 157, 170, 195, 418, 475
- Wechsler Intelligence Scale for Children, 188
evaluation of, 195
validity of, 191
- Weider, H. A., 475
- Weighted score, 179, 206
- Weigl-Goldstein-Scheerer Color Form Sorting Test, 440
- Wellman, B. L., 237
- Welsh, G. S., 475
- Wertheimer, M., 447
- Wesman, A. G., 329
- Wexler, M., 520
- Whipple, G. M., 102
- Willerman, B., 564
- Williams, M., 527
- Windle, C., 526
- Wing, H. D., 340
- Wing Standardized Tests of Musical Intelligence, 337
- Wishner, J., 527
- Wold, J. A., 444
- Wolf, R., 459
- Wolff, H. G., 475
- Wolff, W., 569
- Wood, L., 211
- Word association tests, 547
- Word definitions, 53
- Word fluency, 280
- Word fluency factor, 78
- Work sample, 401
- World War I, 241, 313
- Wright, B. A., 567
- Wrightstone, J. W., 324, 397
- Yale Educational Aptitude Battery, 291
- Yerkes, R. M., 115, 242
- Young, R. A., 525
- Zia Indians, 268
- Zyve, D. L., 358







Form No. 3.

PSY, RES.L-1

**Bureau of Educational & Psychological
Research Library.**

The book is to be returned within
the date stamped last.

19 JUN 1963	25 JUN 1973	
19 JUN 1963	28/9/73	
22 JUN 1963	14.9.73	
15.7.65	3.10.75	
25 JUN 1968		
13 JUN 1969		
20 JUN 1960		
5.1.70		
-6 APR 1972		
20 APR 1972		
7.7.72		
4.2.73		
12.4.73		

151.2
FRE

1579

